

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Volume II

1942

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

A quarterly journal devoted to the development and application of
measures of individual differences.

EDITOR

G. FREDERIC KUDERUnited States Civil Service Commission

ASSOCIATE EDITORS

DOROTHY C. ADKINS.Social Security Board

FORREST A. KINGSBURYUniversity of Chicago

M W RICHARDSONAdjutant General's Office, A. U. S.

BOARD OF COOPERATING EDITORS

RICHARD D. ALLEN
Providence Public Schools

P. J. RULON
Harvard University

JOHN G. DARLEY
University of Minnesota

DAVID SEGEL
U. S. Office of Education

HAROLD A. EDGERTON
Ohio State University

C. L. SILARTILF
Social Security Board

MAX D. ENGELHART
Chicago City Junior Colleges

H. C. TAYLOR
Western Electric Company

E. B. GREENE
Social Security Board

THELMA G. THURSTONE
Chicago Teachers College

J. P. GUILFORD
University of Southern California

HERBERT A. TOOPS
Ohio State University

E. F. LINDQUIST
State University of Iowa

E. G. WILLIAMSON
University of Minnesota

BEN D. WOOD
Columbia University

The journal is open to (1) reports of research on the development and use of tests and measurements in education, government, and industry, (2) descriptions of testing programs being used for various purposes, (3) discussions of problems of measurement in general or in specific fields, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of types of items or improved methods.

INDEX FOR VOLUME II

<i>Aldrich, Margaret Glockler</i>	
AN EXPLORATORY STUDY OF SOCIAL GUIDANCE AT THE COL- LEGE LEVEL	209
<i>Allen, Richard D. and Kline, Lester F.</i>	
EDUCATIONAL REQUIREMENTS AND OCCUPATIONAL LEVELS.	371
<i>Barnes, Melvin W.</i>	
A TECHNIQUE FOR TESTING UNDERSTANDING OF THE VISUAL ARTS	349
<i>Beers, F. S.</i>	
THE EXAMINERS OFFICE OF THE UNIVERSITY SYSTEM OF GEORGIA	233
<i>Berdie, Ralph F.</i>	
AN AID TO STUDENT COUNSELORS	281
<i>Brown, Edwin J.</i>	
SOME OF THE LESS MEASURABLE OUTCOMES OF EDUCATION.	353
<i>Cardall, Alfred J.</i>	
A TEST FOR PRIMARY BUSINESS INTERESTS BASED ON A FUNCTIONAL OCCUPATIONAL CLASSIFICATION.	113
<i>Churchill, Ruth D., Gurtis, Jeanne M., Coombs, Clyde H., and Harrell, Thomas W.</i>	
EFFECT OF ENGINEER SCHOOL TRAINING ON THE SURFACE DEVELOPMENT TEST	279
<i>Cleveland, Earle, Faubian, Richard W., and Harrell, Thomas W.</i>	
APTITUDE TESTS FOR ARMY WEATHER OBSERVER STUDENTS	335
<i>Crissy, William J. E. and Wantman, M. J.</i>	
MEASUREMENT ASPECTS OF THE NATIONAL CLERICAL ABILITY TESTING PROGRAM	37
<i>Evans, Catharine and Wrenn, G. Gilbert</i>	
INTROVERSION-EXTROVERSION AS A FACTOR IN TEACHER- TRAINING	47
<i>Faubian, Richard W., Cleveland, Earle A., and Harrell, Thomas W.</i>	
THE INFLUENCE OF TRAINING ON MECHANICAL APTITUDE TEST SCORES	91
<i>Froehlich, Clifford</i>	
A STUDY OF THE GENIUS VOCATIONAL INVENTORY.	75
<i>Gulford, J. P., Lovell, Constance, and Williams, Ruth M.</i>	
COMPLETELY WEIGHTED VERSUS UNWEIGHTED SCORING IN AN ACHIEVEMENT EXAMINATION.	15
<i>Hahn, Milton E.</i>	
LEVELS OF COMPETENCY IN CHEMISTRY	

<i>Jurgensen, Clifford E.</i>	
A TEST FOR SELECTING AND TRAINING INDUSTRIAL TYPISTS	409
<i>Koran, Sidney W.</i>	
MACHINES IN CIVIL SERVICE TESTING	167
<i>Ligon, Ernest M.</i>	
THE ADMINISTRATION OF GROUP TESTS	387
<i>Lorr, Maurice and Meister, Ralph K.</i>	
THE OPTIMUM USE OF TEST DATA	339
<i>Lurie, Walter A.</i>	
THE CONCEPT OF OCCUPATIONAL ADJUSTMENT	3
<i>McQuitty, John V.</i>	
PROCEDURE FOR HANDLING TESTS AND EXAMINATIONS	153
<i>Mosier, Charles I.</i>	
MEASUREMENT IN RURAL HOUSING; A PRELIMINARY REPORT	139
<i>Oberheim, Grace M.</i>	
THE PREDICTION OF SUCCESS OF STUDENT ASSISTANTS IN COLLEGE LIBRARY WORK	374
<i>Owens, William A. Jr.</i>	
INTRA-INDIVIDUAL DIFFERENCES VERSUS INTER-INDIVIDUAL DIFFERENCES IN MOTOR SKILLS	299
<i>Sabin, Theodore R. and Anderson, Hedwin C.</i>	
A PRELIMINARY STUDY OF THE RELATION OF MEASURED INTEREST PATTERNS AND OCCUPATIONAL DISSATISFACTION	23
<i>Schrammel, H. E.</i>	
THE PURPOSE, ORIGIN, PLAN OF PROCEDURE, AND VALUES OF THE NATION-WIDE EVERY PUPIL SCHOLARSHIP TESTS	401
<i>Sperling, Abraham</i>	
A COMPARISON OF THE HUMAN BEHAVIOR INVENTORY WITH TWO OTHER PERSONALITY INVENTORIES	291
<i>Super, Donald E.</i>	
THE PLACE OF APTITUDE TESTING IN THE PUBLIC SCHOOLS	267
<i>Tussing, Lyle</i>	
AN INVESTIGATION OF THE POSSIBILITIES OF MEASURING PERSONALITY TRAITS WITH THE STRONG VOCATIONAL INTEREST BLANK	59
<i>Watson, Robert I.</i>	
THE RELATIONSHIP OF THE AFFECTIVE TOLERANCE INVENTORY TO OTHER PERSONALITY INVENTORIES	83
<i>Welker, E. L. and Harrell, T. W.</i>	
PREDICTIVE VALUE OF CERTAIN "LAW APTITUDE" TESTS	201
<i>Wood, Ray G.</i>	
THE AIMS, OBJECTIVES, AND OUTCOMES OF THE OHIO TESTING PROGRAM	22

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Volume II

JANUARY, 1942

Number 1

THE CONCEPT OF OCCUPATIONAL ADJUSTMENT	3
<i>Walter A. Lurie</i>	
COMPLETELY WEIGHTED VERSUS UNWEIGHTED SCORING IN AN ACHIEVEMENT EXAMINATION	15
<i>J. P. Guilford, Constance Lowell, and Ruth M. Williams</i>	
A PRELIMINARY STUDY OF THE RELATION OF MEASURED INTER- EST PATTERNS AND OCCUPATIONAL DISSATISFACTION	23
<i>Theodore R. Sarbin and Hedwin C. Anderson</i>	
MEASUREMENT ASPECTS OF THE NATIONAL CLERICAL ABILITY TESTING PROGRAM	37
<i>William J. E. Grissy and M. J. Huntman</i>	
INTROVERSION-EXTROVERSION AS A FACTOR IN TEACHER-TRAINING	47
<i>Catharine Evans and C. Gilbert Wrenn</i>	
AN INVESTIGATION OF THE POSSIBILITIES OF MEASURING PERSON- ALITY TRAITS WITH THE STRONG VOCATIONAL INTEREST BLANK	59
<i>Lyle Tussing</i>	
A STUDY OF THE GENTRY VOCATIONAL INVENTORY	75
<i>Clifford Froehlich</i>	
THE RELATIONSHIP OF THE AFFECTIVE TOLERANCE INVENTORY TO OTHER PERSONALITY INVENTORIES	83
<i>Robert I. Watson</i>	
THE INFLUENCE OF TRAINING ON MECHANICAL ABILITY TEST SCORES	91
<i>Richard W. Faubion, Earle A. Cleveland, and Thomas W. Howell</i>	

Copyright, 1942, by
SCIENCE RESEARCH ASSOCIATES

PRINTED IN THE UNITED STATES OF AMERICA

THE CONCEPT OF OCCUPATIONAL ADJUSTMENT¹

WALTER A. LURIE

Jewish Vocational Service and Employment Center, Chicago

I. THE PROBLEM

THE NEED for criteria of occupational adjustment arises from the attempt to evaluate educational and vocational guidance programs. Many criteria have been proposed, most of them falling into one of the following groups: earnings, job performance, job satisfaction, stability of employment, level of work done, social value of work done, and realization of potentialities.

Several more recent papers have dealt specifically with the weaknesses of studies based on various criteria. Stott (2), in summarizing British experience with a number of the proposed criteria, stressed particularly the sources of unreliability in each of the suggested estimates. Williamson and Bordin (8) have reviewed studies which they are careful to designate as "evaluation of counseling programs" rather than of "adjustment," pointing out structural defects and specific weaknesses in these investigations. Viteles, who had in 1932 suggested accepting "satisfaction and economic efficiency as independent criteria of adjustment in work" (6, p.140), in 1936 proposed a clinical measure, the "dynamic criterion" (7), based upon the extent to which the individual had realized his capacity for vocational success. Williamson and Bordin (8) advocate a "judgment criterion," also a clinical estimate.

These various approaches delimit four possible methods of evaluating vocational adjustment.

¹I wish to acknowledge my gratitude to Dr. Irving Lorge for supplying me with some of the data used in this study. I also wish to thank Mr. S. T. Friedman for assistance with the computations.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

A. The first method is to accept one of the proposed criteria as the essence of occupational adjustment and to define the various degrees of excellence of adjustment in terms of that variable alone. Investigations of vocational adjustment are simply made in terms of the relative job-satisfaction of the individuals, or of their relative earnings, and so forth. The choice of any single criterion is obviously arbitrary.

B The second possible method of evaluating vocational adjustment is to observe two or more of the criterion variables and to combine the ratings into a single score of vocational adjustment. One might, for instance, weight job satisfaction as 5, skill as 2, earnings as 2, job status as 1, and designate the weighted criterion score as a measure of occupational adjustment. Even if a more complex functional relationship is postulated, the arbitrary and inflexible nature of this procedure is immediately apparent.

C. The clinical or judgmental method of evaluating adjustment is in most current use. Williamson and Bordin (8, p.17) describe the "judgment criterion" as one "by means of which the adjustment of the student is *estimated* in terms of his original problems and of the available data, including the part criteria" (i.e., the various separate criteria which have been proposed). In contrast to the first two methods, the clinical method uses all the available data, rather than a restricted set. The method of combining data is not mechanical and inflexible, as in an arithmetical combination. The person making the judgment is expected to take individual circumstances into consideration, in an effort to obtain an integrated representation of the adjustment of each unique personality. By the exercise of clinical insight, when combining data, the judges in effect assign a set of weights characteristic of the individuals judged: earnings are unimportant for John Doe, who comes from a wealthy family; the neurotic Jane Smith will never be more satisfied in any other job than in this one; job status is particularly important for Richard Roe because of his brother-in-law's position in the community; and similar examples. While the skill and intuition of the counselor

THE CONCEPT OF OCCUPATIONAL ADJUSTMENT

obviously play an important role in assigning individuals to places on the scale of vocational adjustment, fair agreement is obtained among judges (9).

There is, however, one basic assumption upon which the validity of the clinical method as an instrument of analysis depends entirely. *It assumes that there is such a psychological entity as occupational adjustment*, that it is therefore possible to define a linear continuum corresponding to excellence of adjustment in various degrees and to locate individuals at definite points on this scale. The clinical method of evaluation shares this fundamental assumption with the first two procedures, the selection of a single criterion and the combination of criteria by formula. It differs only in advocating clinical judgment as the instrument for assigning individuals to points on the scale of adjustment.

It is not my purpose, in questioning the validity of this assumption, to disparage the role of clinical insight in counseling. No one who has attempted to guide individuals in the choice of careers and preparation for them, and later to evaluate the results, can deny that clinical intuition gives more full-bodied and meaningful results than the use of a single criterion or the mechanical combination of criteria. This does not preclude the possibility that the method rests upon a faulty premise. It is always possible to project data of any degree of complexity upon a single axis, either by formula or flexibly, on a case-by-case basis. But this process will necessarily be arbitrary and meaningless if the structure is not, in actual fact, unidimensional. If there is no such linear continuum as occupational adjustment, all three methods which we have considered would be invalidated, the clinical as well as the other two. The superior satisfaction which the counselor receives from the use of the clinical method may reflect an actual lack of precision, which covers up more effectively than the mechanical procedure the incompatibility of data combined in his judgment. The only check upon his estimates, namely, his agreement with other judges, may show merely the extent to which they share his preconceptions.

D. It is therefore advisable to seek a fourth procedure for evaluating occupational adjustment, differing from the first three in its fundamental premise. The nature of such a procedure is clear in a simpler, but fully analogous, situation.

Let us suppose that it is our problem to evaluate the size of human adults, instead of their occupational adjustment. The simplest thing to do would be to define size in terms of, let us say, weight. This would be comparable to identifying occupational adjustment with job satisfaction. But why not choose height, or volume? The choice is arbitrary. A second possibility is to develop a formula combining height and weight. This would require many individual exceptions, because of differences in sex, age, skeletal structure, incidence of crippling accident or disease, and other factors. Would the next proposal then be to substitute a clinical or judgmental combination of all the various factors for the purpose of arranging individuals in order of size? It is more likely that investigators would suspect size to be a variable which cannot be evaluated as such, because no continuum corresponds to the concept. Efforts would be directed first towards defining a set of variables which have something to do with size as popularly conceived and which can be observed and predicted. A further step would be to investigate the dimensionality and structure of the "size" variables. Basic variables—primary factors—would be identified in terms of which all the "part-criteria" of size could be studied in relation to various hereditary and environmental influences.

Now let us reverse the analogy. It is the contention of this paper that occupational adjustment, like size, is a composite variable. Since many part-criteria have already been defined, the next step in evaluating vocational adjustment as popularly conceived is to investigate its dimensionality and structure. Unless this next step is taken, we must work with a concept of occupational adjustment which brings together in a single rating or judgment different factors which have meaning separately but which cancel each other when an effort is made to combine them by formula or by the exercise of clinical insight.

THE CONCEPT OF OCCUPATIONAL ADJUSTMENT

II THE EVIDENCE

The choice among the four proposed methods can be submitted to consideration in the light of evidence. If the evidence shows that there is a single factor common to all proposed criteria which can in themselves be considered meaningful, the concept of vocational adjustment as a psychological entity will have been upheld. Whether we should then use a single criterion, a combination by formula, or a clinical estimate would be a practical problem, depending upon which could be shown to arrange individuals most accurately in order of excellence of vocational adjustment. If, however, several independent factors are demonstrated, it would obviously be wiser to regard these as separate criterion-variables, which must be observed separately and thought of separately as goals of guidance programs. A crucial test of these various approaches can, therefore, be applied by the factor analysis of criterion data.

Table 1 B shows the intercorrelations of five criterion variables (Table 1 A) used in Thorndike's (3) study of prediction of vocational success, for 175 men in the biennium 24-25. Table 1 C shows the distribution of tetrad differences from these correlations. It seems likely, in view of the large deviations from zero, that one factor is not sufficient to account for the intercorrelations of these carefully collected data.

Table 2 B shows the tetrachoric correlation coefficients (1) of twelve criterion and background items (Table 2 A) for 55 female job-applicants born in 1914, whose contact with the Jewish Vocational Service occurred in 1937. Table 2 C gives the centroid factor loadings (4), Table 2 D the distribution of residuals after extraction of three factors, Table 2 E the final factor loadings, and Table 2 F the matrix of transformation by which the centroid factor loadings in Table 2 C were rotated to obtain the final factor loadings as given in Table 2 E (5). Table 2 G shows that the factors are by no means identical; the greatest deviation from orthogonality is less

than 25 degrees, and one plane is orthogonal to both of the others. The factors can be identified only tentatively, in view of the few subjects and the difficulty in obtaining precise information about the clients of an employment agency. Table 2 H lists the high and zero loadings for each factor. Factor I seems to be a reflection of the amount of work experience, which was not controlled; Factor II is primarily a matter of job-satisfaction or job-level; and Factor III is an employability factor. At any rate, it is clear that the criterion vectors do not lie along a single axis.

III. CONCLUSIONS AND DISCUSSION

A It is my belief that the evidence warrants a tentative verdict in favor of the fourth approach, that of discarding the concept of occupational adjustment as a psychological entity and observing separately instead several dimensions of occupational adjustment. While the force of this study may be diminished because it was necessary to use weak criteria and scanty data, and in particular because no clinical estimates could be included, it is certainly not invalidated by these faults. The burden of proof rests now upon those who would evaluate occupational adjustment on a linear scale. They must show that the polydimensionality demonstrated in this study reflects the introduction of irrelevancies and that all meaningful criteria *can* be projected on a linear scale.

B. The present study is not sufficiently comprehensive to identify clearly the dimensions of vocational adjustment. It seems likely, however, that job-satisfaction is closely identified with one factor, and ease of obtaining employment with another. Fruitful investigations into the nature of vocational adjustment can be directed towards further clarification of these factors and of their relationship to the potentialities, background, and aspirations of the individuals.

C. These results, even if they are accepted as conclusive, do not by any means throw into question the efficacy of clinical insight in the counseling process. Recommendations to individuals regarding their life-conduct can be made only by

THE CONCEPT OF OCCUPATIONAL ADJUSTMENT

skilled advisers, never automatically by formula. The clinical method may have a place in evaluation as well as counseling; job satisfaction, for instance, may be better estimated than measured by objective techniques, and this may be true of any other variable which is a psychological entity. Perhaps, however, if these findings are correct, counselors will be able to sharpen their conception of the goals which they are attempting to promote for each individual. When they evaluate informally the results of counseling in a particular case, they may find three or four sentences more informative than an over-all judgment regarding the excellence of occupational adjustment.

D. Finally, it cannot be argued that we are justified in projecting all criterion data on a single axis because it is *important* to have an over-all evaluation of adjustment. What is important is to know the nature, the character of each individual's vocational adjustment. We must study the whole person in action in terms of meaningful variables of the occupational adjustment complex, but we need not place him on a meaningless scale.

IV. SUMMARY

A. It was shown, by logical analysis of various possible methods of evaluating excellence of occupational adjustment, including clinical judgments, that they all assume such estimates or judgments to form a linear continuum.

B. Evidence was presented that only polydimensional models can represent adequately two sets of data, one from Thorndike's major study of vocational success, one derived from clients of the Jewish Vocational Service.

C. Without questioning the efficacy of clinical insight in the guidance process, it was suggested that the concept of occupational adjustment as a psychological entity should be supplanted by a concept of occupational adjustment as a complex of factors which must be observed separately and can be considered separately as goals of guidance programs.

D. Further studies should be directed towards the more precise identification of these factors. This first tentative

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

identification would suggest that one has to do with job-satisfaction and one with ease of obtaining employment, in addition to others not as yet identified.

TABLE 1*

Data from Thorndike's study
175 men in biennium 1924-1925

1 A. <i>Variables</i>						1 C. <i>Tetrad differences***</i>	
						Absolute Value	Frequency
1. Earnings						00	1
2. Level of job						01	2
3. Interest in job						02	0
4. Time unemployed						03	1
5. Change of employer						04	1
						05	0
						06	2
						07	3
						08	0
						09	0
						10	0
						11	1
						12	1
						13	1
						14	0
						15	0
						16	2

1 B. <i>Intercorrelations**</i>					
Var.	1	2	3	4	5
1		45	21	-60	-34
2			32	-14	-18
3				-06	-15
4					40
5					

The standard errors of these coefficients range from .03 to .07.

*Decimal points have been omitted.

**I wish to thank Dr. Irving Lojce for supplying me with this table of intercorrelations.

***Only one of each pair of tetrad differences identical except for sign has been recorded

THE CONCEPT OF OCCUPATIONAL ADJUSTMENT

TABLE 2¹

Data from JVS & EC clients

55 women born in 1914, data as of 1937

2 A. Variables

1. Time of leaving full-time school; 1933 or later recorded as minus (+1% recorded as minus)
2. Number of months employed to 1937; 39 or less recorded as minus (43%)
3. Minimum weekly wages on jobs reported; \$12 or less, minus (38%)
4. Maximum weekly wages reported; \$15 or less, minus (42%)
5. Number of different employers reported; 2 or less, minus (46%)
6. Minimum wage stated to be acceptable, less than \$15, minus (29%)
7. Satisfaction with wages; will not consider wage lower than previous maximum, minus (35%)
8. Waiting time for first job, unemployed until year following time of leaving school, minus (39%)
9. Success of JVS efforts to place in job; no placement, minus (63%)
10. College or specialized training beyond 4 year high school; none, minus (37%)
11. Satisfaction with type of work done; seeking other type of work, minus (13%)
12. Freedom from recorded handicap, including speech defect, language handicap, deformity, etc.; handicap present, minus (13%)

¹Decimal points have been omitted.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 2 (Continued)

2 B. <i>Tetrachoric correlation coefficients</i>												
Var	1	2	3	4	5	6	7	8	9	10	11	12
1		69	-13	34	00	48	-12	-52	-07	-57	25	11
2			56	62	-18	58	08	00	-07	07	64	18
3				34	-16	35	-24	-12	-23	26	59	19
4					01	59	63	07	-10	13	85	09
5						21	07	43	32	19	12	80
6							-41	16	00	-06	12	57
7								40	09	20	70	60
8									20	31	06	65
9										40	-50	10
10											-10	-50
11												47
12												

The standard errors of these correlations range up to about .15.

2 C. *Centroid factor loadings*

Var.	I	II	III	Communality
1	66	-22	-18	52
2	73	36	14	68
3	53	35	32	50
4	49	60	38	75
5	-18	54	-60	68
6	63	33	-43	69
7	-25	42	57	56
8	-30	72	-31	71
9	-43	22	-27	30
10	-44	49	37	57
11	62	62	49	1.00
12	32	43	-86	1.02

2 D. *Distribution of absolute values of third factor residuals*

Absolute Value	Frequency
00-04	11
05-09	13
10-14	10
15-19	8
20-24	13
25-29	4
30-34	5
35-39	0
40-44	2

2 E. *Final factor loadings*

Var.	I	II	III
1	53	-22	-10
2	82	45	00
3	63	52	-11
4	69	76	00
5	01	07	82
6	69	08	44
7	-06	64	-14
8	-02	38	73
9	-33	-02	41
10	-22	57	09
11	82	85	-09
12	43	-11	88

2 F. *Matrix of transformation*
(Λ_n)

	I	II	III
I	930	100	-156
II	368	812	618
III	031	576	-770

2 G. *Direction cosines of normals*
($\Lambda'_n \Lambda_n$)

	I	II	III
I		410	058
II			043
III			

THE CONCEPT OF OCCUPATIONAL ADJUSTMENT

TABLE 2 (Continued)

2 H. *Tentative identification of factors*

Factor I.

A. *Variables with loadings over .40*

2. Number of months employed...	82
11. Satisfaction with type of work	82
6. Minimum wage acceptable.	69
4. Maximum weekly wages.	69
3. Minimum weekly wages ..	63
1. Time of leaving school.	53
12. Freedom from handicap.	43

B. *Variables with loadings under .15*

5. Number of different employers.	01
8. Waiting time for first job.	02
7. Satisfaction with wages.	-06

C. *Tentative identification*

Experience

Factor II.

A. *Variables with loadings over .40*

11. Satisfaction with type of work	85
4. Maximum weekly wages ..	76
7. Satisfaction with wages.	64
10. Special training	57
3. Minimum weekly wages.	52
2. Number of months employed.	45

B. *Variables with loadings under .15*

9. JVS placements	02
5. Number of different employers.	07
6. Minimum wage acceptable.	08
12. Freedom from handicap.	-11

C. *Tentative identification*

Job level or job satisfaction

Factor III.

A. *Variables with loadings over .40*

12. Freedom from handicap.	88
5. Number of different employers.	82
8. Waiting time for first job.	73
6. Minimum wage acceptable.	44
9. JVS placements	41

B. *Variables with loadings under .15*

2. Number of months employed.	00
4. Maximum weekly wages.	00
10. Special training	09
11. Satisfaction with type of work	-09
1. Time of leaving school.	-10
3. Minimum weekly wages.	-11
7. Satisfaction with wages.	-14

C. *Tentative identification*

Ease in finding employment

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

REFERENCES

1. Chesire, L., Saffir, M., and Thurstone, L. L. *Computing Diagrams for the Tetrachoric Correlation Coefficient*. Chicago: University of Chicago Press, 1933.
2. Stott, M. B. "Occupational Success." *Occupational Psychology* (London), XIII (1939). 126-140.
3. Thorndike, E. L., et al. *Prediction of Vocational Success*. New York: The Commonwealth Fund, 1934.
4. Thurstone, L. L. *The Vectors of Mind*. Chicago: University of Chicago Press, 1935.
5. Thurstone, L. L. "A New Rotational Method in Factor Analysis." *Psychometrika*, III (1938). 199-218.
6. Viteles, M. S. *Industrial Psychology*. New York: W. W. Norton and Co., 1932.
7. Viteles, M. S. "A Dynamic Criterion." *Occupations*, XIV (1936), 962-967.
8. Williamson, E. G. and Bordin, E. S. "The Evaluation of Vocational and Educational Counseling: A Critique of the Methodology of Experiments." *Educational and Psychological Measurement*, I (1941), 5-24.
9. Williamson, E. G. and Bordin, E. S. "A Statistical Evaluation of Clinical Counseling." *Educational and Psychological Measurement*, I (1941), 117-132.

COMPLETELY WEIGHTED VERSUS UNWEIGHTED SCORING IN AN ACHIEVEMENT EXAMINATION

J. P. GUILFORD, CONSTANCE LOVELL, and RUTH M. WILLIAMS
University of Southern California

IN A PREVIOUS REPORT,¹ the senior author presented the derivation of a scoring weight for differential weighting of responses to test items. The formula for the weight, in a form recommended for practice, is as follows:

$$W = 4 + \frac{p_u - p_l}{pq}$$

in which W = the scoring weight,

p_u = the proportion of an upper (or otherwise specified criterion sub-group) reacting in a defined manner,

p_l = the similar proportion in a lower (different) criterion sub-group,

p = the proportion of the two sub-groups combined responding in this manner, and

$q = 1 - p$.

Such a weight has heretofore been most widely employed in connection with personality tests of the type of the *Strong Vocational Interest Blanks*. In the following study the weight was used in connection with an objectively-scored multiple-choice achievement examination. In this kind of test we can consider the probability of a specified response being made (p) or not being made (q), for a group as a whole, and

¹J. P. Guilford, "A Simple Scoring Weight for Test Items and Its Reliability," *Psychometrika*, VI (1941), 367-374.

also the probabilities of the same response being made within two separated groups. Our main problem was to determine whether an examination with completely weighted scoring of this kind yields any more highly reliable and valid scores than the same examination yields with unweighted scoring. A subsidiary problem was to determine whether the length of examination has any bearing upon the effect of weighted versus unweighted scoring. By "completely weighted" we mean that *every* response, whether considered right or wrong, is given a weight in proportion to its predictive significance. This procedure is in contrast to ordinary differential weighting where only correct responses are weighted in proportion to their diagnostic value. By "unweighted scoring" we mean that items are given weights of 0 or 1 — 0 if a wrong response of any kind is given and 1 if the correct response is given.

As our experimental material we used the results of a final examination in a course on "Problems of Human Behavior."² The test was composed of 201 multiple-choice items and 107 true-false items. These items had been analyzed for validity in previous use; therefore, we could expect to find an unusually large number of diagnostic items both when scoring was weighted and when it was not. Instructions to guess or not to guess were not stated on the test blanks, but the usual correction formula had been applied to allow for guessing. There were extremely few omissions. These facts are mentioned because we used the total examination score as our criterion of achievement in the course.

Our study was confined to the first 100 consecutive items in the examination, skipping one item which had an unusual number of omissions and which had been excluded from the scoring of the total examination. All but 8 of the items proved to have phi coefficients of .14 or larger (in other words, significant correlations)³ and all but 19 had very significant phi

²We are indebted to Dr Neil Warren for the opportunity to use this material

³J. P. Guilford, "The Phi Coefficient and Chi Square as Indices of Item Validity," *Psychometrika*, VI (1941), 11-19.

coefficients (greater than .18). The total range of phi coefficients for the 100 items was from -06 to $+.48$, with a median of $+.28$.

Three hundred test papers, selected at random, were used in this investigation. The papers were re-scored in order to be absolutely sure about total, or criterion, scores. The upper and lower criterion sub-groups were composed of the 100 highest and the 100 lowest ranking students in the list of 300. The proportion of each group responding in each of the four ways to every item was determined. The scoring weight for each particular response was read from a graphic chart.¹ For example, one item read: "Genius and feeble-mindedness are (1) points on a normal distribution curve; (2) points on a bimodal distribution; (3) points on a multimodal distribution; (4) in separate distributions." The scoring weights for responses 1 to 4 inclusive were: 6, 3, 4, and 2, respectively.

A logical reason for expecting improved reliability and validity from weighted scoring is now more apparent. Not all the wrong responses are of equal diagnostic value, an assumption that is implicitly made in unweighted scoring. It is apparently worse for a student to err by choosing answer 4 and less serious for him to choose answer 3. In only 15 items out of the 100 did the three wrong answers prove to have equal weight. No weights exceeded 6 points nor fell below 2 points on a scale that extended from 0 to 8. This range was to be expected from the moderate and small sizes of the phi coefficients found for correct responses, as previously mentioned.

The factor of length of test was investigated roughly by selecting two shorter examinations composed of the first 20 and the first 50 items out of the 100. Each of the three tests of different length was scored in both halves (odds and evens) and in total, with and without scoring weights. For this purpose, 100 papers were selected out of the original list of 300, by taking every third paper when the 300 were in rank order for total scores. The correlations to be mentioned next are based upon these 100 papers.

¹See reference in footnote 1.

TABLE 1
RELIABILITY AND VALIDITY COEFFICIENTS

Length of Test	Reliability		Validity	
	Weighted Scoring	Unweighted Scoring	Weighted Scoring	Unweighted Scoring
20 items667	.649	.817	.793
50 items860	.844	.892	.901
100 items922	.899	.900	.924

The reliability coefficients were estimated by the Spearman-Brown formula in each case. The correlation of each short test with the total (criterion) score was computed. The reliability and validity coefficients are summarized in Table 1. Here it is obvious that the weighted scoring yielded a scant average gain of .02 in the reliability coefficients. This trifling gain is consistent, but seems to be insignificant. In validity the weighted scoring yielded a gain of about .02 in the shortest test and a like amount of loss in the longest test, neither of these changes being significant.

It is well to consider possible special reasons for the failure to obtain increased reliability and validity here. As indicated above, the phi coefficients between items and criterion scores were generally low and the range of differential weights was relatively small. It might be that in other examinations which include items with weights extending closer to 0 and 8 there would be an appreciable gain from differential weighting. On the other hand, 32 of our 100 items had weights ranging from 2 to 6, 21 more had weights ranging from 2 to 5 or from 3 to 6; and 45 had weights ranging from 3 to 5, to be contrasted with an unweighted range of 0 to 1.

One important source of gain to be expected from complete differential weighting comes from the variations in weights among the wrong responses, which in unweighted scoring or in partially weighted scoring are all given the same value of zero. Gaps of more than one point between the correct response on the one hand and all the wrong responses on the other simply magnify numerically the variability among individuals' total scores, except that the more diagnostic items are then allowed to contribute relatively more to the total

WEIGHTED VERSUS UNWEIGHTED SCORING

variability than do the less diagnostic ones. When there is a spread of the weights among the wrong responses, there should be more refinement in providing effectual variability among total scores. The weights for wrong responses among our 100 items showed very narrow ranges. 15 items had equal weights for wrong responses, 68 had differences not exceeding one, and 17 had differences not exceeding two points. While these differences are small, it would seem that their effects should have been felt in scoring.

Another possible factor in the failure of weighting is that the criterion scores were derived from unweighted scoring. Correlations of either weighted or unweighted scores for the shorter tests of 20, 50, and 100 items and total scores are in a sense spurious in that we are correlating part with whole. This factor might have favored higher correlations, especially for the unweighted scoring of the parts.

In using total scores as our criterion of achievement here, we have assumed up to this point that, though the part-whole correlations are spurious, they are equally so for both types of scoring. One evidence that this assumption may not be sound is the fact that the shorter the part-test, the relatively greater is the advantage of weighted over unweighted scoring. (See Table 1) On the other hand, it may be characteristic of short tests *per se* to gain relatively more from weighted scoring. Other evidence is found among the reliability coefficients. Here we have unweighted scoring correlated with unweighted scoring and weighted scoring correlated with weighted scoring. The correlations indicate that regardless of the length of the test within the range of 20 to 100 items, there is about the same slight advantage for weighted scoring. Although changes in validity do not always parallel changes in reliability, it would seem that if the evidence of reliability here is dependable, the systematic variations in validity may be due to the relatively greater spuriousness for unweighted scoring in the longer tests (since they are greater parts of the total and are similarly scored) rather than to any greater relative gain in validity of short tests by weighted scoring. The evidence is

entirely too meager, however, for us to draw any final conclusions on this point.

In view of the uncertainty introduced by the factor of spuriousness of correlation, it would have been interesting to see what would have happened with an outside criterion. Some idea of the extent of spuriousness can be obtained, without taking the trouble to score the tests minus the 20, or 50, or 100 items used in the experimental tests, by applying a formula to estimate the amount of correlation between part and the whole from which the effects of the part are eliminated.⁵ For the 50-item test, with unweighted scoring, for example, this estimated correlation is .862, which may be taken as an indication of the correlation between the 50-item test and an outside criterion of about 250 items. The correlation of the same test with a homogeneous test of 300 items (the approximate length of the total examination) is estimated by formula to be .865.⁶ The amount of spuriousness is then indicated by the difference between .865 and .901. We cannot similarly estimate the amount of spuriousness in the correlation of the weighted scoring in the 50-item test since it is not a simple part-whole relationship. But had the spuriousness in this case been zero, the validity coefficient of .892 is not quite .03 higher than that of the estimated unweighted scoring without its spurious element. It is doubtful, therefore, whether the hypothesis of greater part-whole spuriousness attributed to the unweighted scoring is sufficient to account for the failure of weighted scoring to exhibit superior validity coefficients.

Had all the items in our tests been significantly correlated with the criterion, a difference in favor of weighted scoring might have resulted. Therefore, we selected for comparison the 50 items of highest diagnostic value. The reliability coefficients were then .874 and .873 for weighted and unweighted scoring, respectively, and the corresponding validity coefficients

⁵C. C. Peters and W. R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases* (New York: McGraw-Hill Book Company, 1940), p. 217.

⁶J. P. Guilford, *Psychometric Methods* (New York: McGraw Hill Book Company, 1936), p. 422.

WEIGHTED VERSUS UNWEIGHTED SCORING

were .900 and .904. These coefficients were insignificantly better than the ones derived from the 50-item test in which the items were taken at random. The inference might be that the items used in all our experimental tests were at a sufficiently high level of diagnostic value, taking them collectively, that weighted scoring was of no consequence.

Our general conclusion is that our logically defensible system of completely weighted scoring did not yield an appreciable gain in either reliability or validity in achievement examinations of from 20 to 100 items. While the result is negative so far as the improvement of test technique is concerned, it is useful to know that the customary unweighted scoring, which takes distinctly less time and effort, gives about as reliable and valid results as differential weights afford.⁷ Although this result may not be generalized to all weighting methods and to all kinds of tests, it does suggest the possibility of satisfactory scoring without weighting in places where we now attempt to extract the utmost validity by the use of differential weights. With the increased use of machine scoring, where differential weighting becomes a serious practical problem, it may be well in any case to consider the efficiency of weights 0 and 1 before recommending a system of differential scoring weights.

⁷This general outcome is in line with a conclusion reached on rational grounds by M. W. Richardson, in Paul Horst's *The Prediction of Personal Adjustment* (New York: Social Science Research Council, 1941), 379-401.

A PRELIMINARY STUDY OF THE RELATION OF MEASURED INTEREST PATTERNS AND OCCUPATIONAL DISSATISFACTION¹

THEODORE R. SARBIN

University of Minnesota

and

HEDWIN C. ANDERSON

Minnesota Division of Vocational Rehabilitation

THAT occupational dissatisfaction is associated with a lack of interest typical of successful men in a particular job is a generally accepted hypothesis. This is of special concern to psychologists, particularly if the hypothesis can be verified and predictions of occupational satisfaction made on the basis of interest measurement. In order to test the hypothesis two kinds of data must be analyzed: (1) evidence of occupational dissatisfaction and (2) measures of vocational interest.

Although rating scales for determining job satisfaction have been developed by Hoppock (4) and others, they are difficult to use in a clinic where the clients or patients form a heterogeneous population. They come from many different walks of life, and there are seldom more than a few individuals who are employed by the same organization. Job satisfaction can be described by Hoppock's definition as "any combination of psychological, physiological or environmental circumstances that causes a person truthfully to say 'I am satisfied with my job.' " (4:47)

¹This study is one of a series of studies in process on clinical problems of interest measurement at the University of Minnesota Testing Bureau.

In commercial and industrial organizations, a psychologist may experience difficulty in persuading workers "truthfully" to state their feelings about their work because they are afraid of losing their jobs. During periods of widespread unemployment, especially, an individual may express satisfaction with his job merely because it is a job. By dealing with groups of workers and guaranteeing anonymity by the use of unsigned questionnaires, a psychologist may gather group data, but the anonymity may prevent his relating these data to such variables as personality traits or interests in subsequent clinical study.

The clinical situation in which the present data were gathered gives greater assurance of meeting Hoppock's qualification regarding the truthfulness of the clients' responses. In the first place, all the subjects came to the University of Minnesota Testing Bureau voluntarily. They had heard of the Bureau through friends or business associates. They recognized the Bureau as a disinterested organization which each year assists a small number of out-of-school adults with problems of vocational adjustment. Secondly, they paid a special fee for the service. This fact presumably predisposed them to tell the truth about their occupational experiences. Finally, if a client had difficulty in expressing himself, a trained clinical interviewer assisted him to say the things that he could not or would not otherwise have said. It is reasonable to assume from these three facts that expressions of occupational dissatisfaction were truthful expressions. Having found a usable index of occupational satisfaction, the next step was to find a measure of vocational interest.

According to a recent poll (1), the most widely used measure of vocational interest is the *Strong Vocational Interest Blank*. This instrument is based upon this fundamental assumption:

"If a man likes to do the things which men like who are successful in a given occupation and dislikes to do the things which these same men dislike to do, he will feel at home in that occupational environment. Seemingly, also, he should be

more effective there than somewhere else because he will be engaged, in the main, in work he likes" (6)

The *Strong Vocational Interest Blank* was standardized upon people who were purportedly successful in their occupations. Strong's criteria of occupational success include the following: length of experience in an occupation, annual income, level of education, certification of membership in professional society, and selection by so-called competent authorities. These were used singly or in various combinations.

This is not the place to list the description of Strong's criterion groups but as an illustration we take three occupations. The samples of successful men have all been engaged in the respective occupations for at least the three previous years, and none is over 60 years of age.

Accountant: "Includes 160 general accountants, 54 cost accountants, 65 auditors, and 66 comptrollers and treasurers. Average age equals 37.4 years; education equals 12.3 grade."

Office Worker: "Includes 214 office clerks, bookkeepers and stenographers; 92 office managers; and 200 credit managers. Average age equals 33.2 years; education equals 11.5 grade."

Physician: "Graduates of Yale and Stanford Medical School. Includes 252 physicians and 75 surgeons (no difference of interest between them) 253 are from California, 47 from Connecticut and 9 from New York; the remaining are scattered. Average age equals 40.9 years; education equals 18.5 grade."

In interpreting the results on the *Strong Vocational Interest Blank*, then, it is always necessary to think of the criterion groups which served as the norm, as well as the percentages of the group included under each grade. Thus if an individual scored A on the key for physicians it means that he made a score in the range of the top 69 per cent of the physicians who made up the norm group. A score of B falls in the range of the next 29 per cent; and a score of C falls in the range of the lowest two per cent of the criterion group (3).

For purposes of the present analysis, 100 cases were

selected from the files of the University of Minnesota Testing Bureau for the period 1937 to 1940 on the basis of completeness of data. This sample contained 76 men and 24 women. The cases were so-called "non-college adults", individuals who are accepted by the University Testing Bureau for research and clinical purposes. Those who had gross physical abnormalities, such as paralysis, spasticity, and deafness were not included in this selection of cases. Only individuals who were 25 years of age or more were included. Above this limiting age individuals usually have had some opportunity to establish a work history. The mean age for men was 31.5 with a standard deviation of 5.7 and a range of 25-53; for women, 30.4 with a standard deviation of 7.7 and a range of 25-44.

The educational level of this group appears to be higher than that of the general population. For men, the mean grade completed was 13.8, S.D. 2.5; for women, 14.5, S.D. 2.3.

The occupational status of this group was also higher than that of the general population. According to the *Minnesota Occupational Rating Scales*, 72 per cent rated in the top three categories. In the general population only 22 per cent fall into these three categories.²

From an analysis of these data, the following hypothesis can be tested:

Adults who express dissatisfaction with their current³ occupations show no primary pattern of interest, as measured by the *Strong Vocational Interest Blank*, for the group of occupations in which their current occupation belongs.

The *Strong Vocational Interest Blank* was first analyzed in order to determine the primary pattern of interest. Darley's

²Occupational Class I: professional; Occupational Class II: semi-professional and managerial, Occupational Class III: clerical, skilled trades, retail business. F. L. Goodenough and J. E. Anderson, *Experimental Child Psychology*, (New York: Appleton-Century, 1931), pp. 501-12.

³The data contained herein are concerned with present occupation (or most recent occupation for four cases unemployed at the time of counseling). Data were available on modal occupations but since these coincided with present occupations in over 90 per cent of the cases, the data were not analyzed in terms of the modal occupations.

MEASURED INTEREST PATTERNS

scheme of determining the presence and intensity of patterns of interest on the Strong Blank was utilized (3). A primary pattern is defined as a preponderance of A and B+ scores within the occupations making up a group of factors as revealed by existing factor analysis studies. To illustrate: the verbal or linguistic interest type, Group X on the Strong test, is made up of the following typical occupational titles—advertising man, editor, lawyer. If a client had scores of A, B+, A respectively, on these three keys, he would be considered to have a primary pattern of interest in this group of occupations. If his scores were B+, B, B, he would be rated as having a secondary pattern of interests. A tertiary pattern of interests is defined as a majority of B and B—scores on the keys within any factor or group. In the present study the number of cases was too small to be treated in terms of Darley's fourfold classification. primary, secondary, tertiary, and no pattern. Instead, we considered only two categories:

- (a) Presence of primary pattern (this means presence of primary pattern as defined above in the group which embraces the client's present or most recent occupation).

e.g. Client's present occupation: Automobile Sales man;

Scores on Strong Blank Group IX	
Real Estate Salesman	B+
Life Insurance Salesman	A
Sales Manager	A

- (b) Absence of primary pattern (this means absence of primary pattern in the group which embraces the client's present or most recent occupation)

e.g. Client's present occupation: Lawyer;

Scores on Strong Blank Group X	
Advertising man	B—
Lawyer	C
Editor	C

Each case was classified according to the client's stated complaint. The following categories were used:

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

- (a) Dissatisfied with occupational field.
- (b) Dissatisfied with present job.
- (c) Dissatisfied with present job only because of future prospects.
- (d) No specifically stated dissatisfaction, but seeks vocational and/or educational information or advice.

Each case was also classified according to the clinician's diagnosis, using five broad classifications:

- (a) *Inappropriate vocational choice*: e.g., "He does not have the interests of salesmen." "Has never been interested in mechanical work." "Working with teachers not congenial to this man's values, attitudes, and ideals."
- (b) *Primary personality disorders*: e.g., "social maladjustment", "unhappily married", "neurotic tendencies."
- (c) *Insufficient education or training*: e.g., "lacks stenographic skills to compete with co-workers", "lacks sufficient graduate training to get into junior college teaching", "lacks skills in cost estimating which are required for promotion and increased pay."
- (d) *Inappropriate job placement*: e.g., "clerical skills not being used", "stenographic skills are not adequate", "truck-driving satisfactory, but would be happier if he had a run closer to home and family", "selling satisfactory, but his product is inappropriate."
- (e) *Other*: This includes a small number of which three were characterized as "no problem", the rest as financial, health, or unclassified.

These diagnostic illustrations are stated as single entities. This is somewhat misleading. For many of the cases a multiple diagnosis was made. For example, 28 per cent of the cases of *inappropriate vocational choice* also exhibited neurotic symptoms and mild personality disorders. The data, however, were treated in terms of the diagnosis that was considered by the clinician to be the most significant one.

A third kind of classification was made to determine the

MEASURED INTEREST PATTERNS

frequency of types of treatment or recommendations. These fall under the following headings:

- (a) *Placement advice.* e.g., "Since your interests and abilities fit the picture of successful salespeople, I would recommend that you register with the X and Y employment agencies." "You should seek employment in a more technical field than your present occupation." "You are faced with two alternatives: taking over your father's business, or continuing as an engineer. Your interests and personality traits would suggest that you would be happier as an engineer than as a business man."
- (b) *Additional training recommended:* e.g., "A University Extension Course in Cost Estimating seems indicated." "In order to prepare for the position in mind, you will have to return to college for two years of graduate work." "In order to capitalize on your assets and interest, you should obtain the necessary skills at such a school as The Blank Industrial Training Institute."
- (c) *Psychotherapy:* e.g., recreational therapy, catharsis, helping client to gain insight into family or other conflict situation, suggestive therapy, group therapy, relationship therapy, and so on.
- (d) *Referral to psychiatrist:* Obvious psychiatric problems.
- (e) *No advice or recommendations.*

The results of the analyses are summarized in the three tables. Table 1 shows the clinicians' diagnoses for 76 male clients and how they are related to the presence or absence of primary patterns of interest and also to clients' complaints. Table 2 shows the same data for 24 female clients. Table 3 summarizes the treatment techniques as related to diagnoses for the 100 cases.

Table 1 reveals one fact quite clearly: most adult males who complain of occupational dissatisfaction show no primary pattern of interest in the group of occupations which embraces their present occupation. Sixty-two of the 76 men (82 per

EDUCATIONAL AND PSYCHOLOGICAL MISALIGNMENT

TABLE 1

CLIENTS' STATEMENT OF PROBLEM AND CLINICIANS' DIAGNOSES IN TERMS OF PRESENCE OR ABSENCE
PRIMARY PATTERN OF INTEREST IN CURRENT OCCUPATION

(N=76 Men)

Client's Statement	Clinician's Diagnosis	Inappropriate vocational choice		Primary personality disorders		Insufficient training		Inappropriate job placement		Other		Total
		PRIMARY PATTERN		PRIMARY PATTERN		PRIMARY PATTERN		PRIMARY PATTERN		PRIMARY PATTERN		
		Pres- ent	Abs- ent	Pres- ent	Abs- ent	Pres- ent	Abs- ent	Pres- ent	Abs- ent	Pres- ent	Abs- ent	
Dissatisfaction with occupa- tional field		0	18	3	6	0	1	4	3	0	1	3
Dissatisfaction with specific job		0	3	3	0	1	0	0	1	0	0	4
Dissatisfaction with future of job		0	0	0	1	1	0	0	7	0	0	1
No specifically stated dissatis- faction		2	12	2	7	0	2	1	1	1	0	6
Totals		2	33	8	14	2	3	5	1	1	7	14

TABLE 2

CLIENTS' STATEMENT OF PROBLEM AND CLINICIANS' DIAGNOSES IN TERMS OF PRESENCE OR ABSENCE
PRIMARY PATTERN OF INTEREST IN CURRENT OCCUPATION

(N=24 Women)

Client's Statement	Clinician's Diagnosis	Inappro- priate vocational choice		Primary personality disorders		Inappro- priate job placement		Other		Total
		PRIMARY PATTERN		PRIMARY PATTERN		PRIMARY PATTERN		PRIMARY PATTERN		
		Pres- ent	Abs- ent	Pres- ent	Abs- ent	Pres- ent	Abs- ent	Pres- ent	Abs- ent	
Dissatisfaction with occupational field,		0	4	1	0	0	0	0	0	1
Dissatisfaction with specific job		0	1	0	1	1	1	0	1	1
No specifically stated dissatisfaction		1	0	1	4	0	1	4	1	8
Totals		1	5	4	5	1	2	4	2	10

TABLE 3

ANALYSIS OF TREATMENT TECHNIQUES IN TERMS OF CLINICIAN'S DIAGNOSIS

(N=76 Men, 24 Women)

Treatment Used	Clinician's Diagnosis	Inappropriate vocational choice		Primary personality disorders		Insufficient education		Inappropriate job placement		Other		Total
		PRIMARY PATTERN		PRIMARY PATTERN		PRIMARY PATTERN		PRIMARY PATTERN		PRIMARY PATTERN		
		Men	Wom- en	Men	Wom- en	Men	Wom- en	Men	Wom- en	Men	Wom- en	
Placement advice		11	2	4	3	1	0	4	3	1	0	21
Recommended additional train- ing		19	4	3	0	4	0	2	0	1	2	29
Psychotherapy		3	0	11	3	0	0	0	0	0	2	14
Referral to psychiatrist		0	0	3	3	0	0	0	0	1	0	4
No advice		2	0	1	0	0	0	0	0	3	2	6
Totals		35	6	22	9	5	0	6	3	8	6	76

MEASURED INTEREST PATTERNS

cent) fall into this category. When we consider the women who came to the Testing Bureau, the association is not so clear cut. (See Table 2) Fourteen of the 24 women clients (58 per cent), had no primary pattern in the occupation in which they were employed. When we consider the 10 women who actually *expressed* dissatisfaction with their work (items 1 and 2 in Table 2), we see that eight (80 per cent) had no primary pattern in the group of occupations which embraced their present employment. These equivocal results in the case of the women may be attributed to the inadequacy of the Strong Blank for Women, to the smaller number of cases, or to the generally accepted statement of sex differences in intensity and variety of interests.

Where the men actually expressed dissatisfaction, 70 per cent referred to the occupational field rather than to the particular job in which they were employed at the time of the interview (Items 1 and 3 versus totals of items 1-2-3). For example, one junior high school teacher said: "It isn't my job at the Blank Junior High School that I don't like. As teaching jobs go, it's a good one. It's the idea of spending the rest of my life as a teacher that is my bogey." A small number did express dissatisfaction with their present jobs, but considered the occupational field in which the job was located as a desirable one. Only five individuals expressed satisfaction with the field of the occupation and with the present job, but were concerned over the future prospects of the job.

Analysis of the clinicians' diagnoses reveals first, that the diagnosis *inappropriate vocational choice* is the most frequently-made diagnosis. Of the 35 men and six women who were diagnosed in this way, 33 men and five women did not show a primary pattern of interest on the Strong Blank in the group of occupations which included their present occupations. Reading of the case notes indicated that when the interest test data were out of line with the present vocation, the clinician almost invariably recorded the diagnosis as *inappro-*

priate vocational choice, and was unable to find any other diagnostic description more appropriate to the facts.

What diagnoses are made when the interest pattern *agrees* with the present employment? In 24 cases (14 men, 10 women) the interest test showed a primary pattern which coincided with the present job of the individual. Fourteen of these (6 men, 8 women) expressed no specific dissatisfaction. Of the 24 clients in this group, half were diagnosed as having primary personality disorders. Of the remaining 12, the diagnoses were about evenly distributed among the other categories. The following hypothesis may be formulated from these data: a person may have the vocational interests of people successfully engaged in his present occupation, but deep-seated personality disorders may otherwise interfere with his occupational adjustment.

Table 3 represents the types of treatment commonly employed by Testing Bureau clinicians. It is beyond the scope of this paper to deal with the evaluation of the various kinds of treatment. The treatment techniques most frequently used were: *placement advice* and *recommended additional training*. These were used primarily where the diagnosis was *inappropriate vocational choice*. As indicated before, these diagnoses (and therefore the treatments) were based on data from the *Strong Vocational Interest Blank*. *Psychotherapy* and *referral to psychiatrist* were employed in most cases which were diagnosed as primary personality difficulties.

It seems quite clear that these data allow us to test only a limited hypothesis. Actually, we are using as reference points Strong's original criterion groups. The conclusions, therefore, can only be stated tentatively until more extensive samples are utilized. If we had selected 100 adults at random among individuals who had not come to the Testing Bureau, how many would have shown primary patterns of interest which coincided with their present occupation? What proportion would have been dissatisfied with their work? What

proportion would have adjusted to such dissatisfaction without the help of an outside agency?

Partial answers to these questions are implied in a recent monograph by Darley (3). He says that individuals who continue in occupations which are at variance with their interest pattern may:

- "(1) Develop socially acceptable and compensatory hobbies;
- (2) Develop personality conflicts at home or on the job, but still keep on the job;
- (3) Re-define the specific job duties more in line with the activities of the primary interest type . . . ;
- (4) Establish a sufficiently poor work record to be only marginally employable (without promotion) or to be separated from the job . . . "

It is not improbable that this sampling is, in the main, composed of individuals who, while they may react in the alternative ways indicated by Darley, also seek the help of an available outside agency in finding an adjustment when they experience dissatisfaction.

To answer the questions raised in this discussion, crucial experiments must be carried out. Until such research is prosecuted, conclusions from these data must be made with caution. The data seem to justify this conclusion: occupational dissatisfaction is associated with a lack of primary interest in the current occupation. What explanations may be offered? Two alternatives immediately suggest themselves:

- (1) A person's interests are temporally stable; they are relatively crystallized *prior* to entry into the occupational world; when the occupational activities and the interests are at variance, dissatisfaction results. The dissatisfaction is a consequent or a resultant of a fixed personality interacting within an occupational milieu.
- (2) A person's interests are temporally not stable; they are flexible and subject to change *subsequent* to entry into the occupational world; they may change as a result of lack of success, environmental factors, or more fundamental personality traits in interaction. The dissatisfaction is antecedent to, or coincident with, changes from a primary pattern of interests to no

primary pattern of interests in the present occupational group.

To know which of these alternative explanations is correct is important for clinicians who are approached for assistance by vocationally-dissatisfied clients, and also by clients who are about to make a vocational choice. If interests are fixed by the time an individual is ready to seek employment, and if dissatisfaction will result if the client enters an occupation outside his interest type, then the clinician will advise him to seek employment in certain restricted areas. If, on the other hand, measured interests and satisfactions are the product of successful achievement, then the clinician will advise clients to seek employment where the greatest possibilities for success are to be found in terms of the clients' abilities and also employment opportunities. Extensive longitudinal studies will determine which of these alternatives, if either, is correct. According to Darley's review of the literature, the first alternative seems more in line with available evidence (3).

A word is in order relevant to the psychological processes which are represented by the *Strong Vocational Interest Blank*. Cartel, in using this instrument, suggests that patterns of interest "become closely identified with the self." Further, "the pattern of interests is in the nature of a set of values . . ." (2). In this connection Sarbin and Berdie have demonstrated that certain relations exist between values as measured by the Allport-Vernon Scale and interests as measured by the Strong Blank (5). It is postulated that the summation of the 400 preferences on the Strong Blank reveals—at least in part—a cross-section of what the individual would like to be; in short, a person's ideal conception of the self. The Freudian expression, *ego-ideal*, carries approximately the same meaning.

Expressed occupational dissatisfaction, then, may be a resultant of the conflict between the *ego-ideal* and the occupational milieu or reality in which the individual applies this conception of the self. When the reality-situation is such that the individual's idealistic self-conception is tested and verified,

no conflict or dissatisfaction ensues. When reality prevents the testing and verification of one's ego-ideal, we find expressions of occupational dissatisfaction.

This interpretation throws no further light on the previously-posed problem as to which of the two alternative explanations is the appropriate one. The problem is merely restated in this form: is one's conception of the self (ego-ideal) a stable phenomenon or is it a variable one? Does it change with each variation in reality, with success and failure experiences? Further experimental work will illuminate some of these dark corners.

Summary

Adults who complain of occupational dissatisfaction show, in general, measured interest patterns which are not congruent with their present or modal occupations. *If* vocational interests are stable temporally, and *if* they have the dynamic character usually attributed to them, we may expect a high incidence of occupational maladjustment when individuals enter occupations for which they do not have the appropriate interests at the time of entry.

REFERENCES

1. Beane, B., Carroll, J., and Habbe, S. "The Beane Poll of Favored Psychological Tests", *Journal of Applied Psychology*, XXIV, (1940), 347-352.
2. Carter, H. D. "The Development of Vocational Attitudes", *Journal of Consulting Psychology*, IV, (1940), 185-191.
3. Darley, John G. *Clinical Aspects and Interpretation of the Strong Vocational Interest Blank*. New York: The Psychological Corporation, 1941, 71 pp.
4. Hoppock, Robert. *Job Satisfaction*. New York: Harper and Brothers, 1935.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

5. Saabin, Theodore R. and Berdie, Ralph. "The Relation of Measured Interests to the Allport-Vernon Study of Values", *Journal of Applied Psychology*, XXIV, (1940).
6. Strong, Edward K., Jr. *Manual for Vocational Interest Blank for Men*. Stanford University Press, 1940.

MEASUREMENT ASPECTS OF THE NATIONAL CLERICAL ABILITY TESTING PROGRAM

WILLIAM J. E. CRISSY

Cooperative Test Service

and

M. J. WANFMAN

University of Rochester

THE PURPOSE of the present article is threefold: to discuss the measurement procedures employed by the committee responsible for the National Clerical Ability Testing Program; to cite some of the measurement problems that confront the committee; and to suggest possible improvements in the procedures and possible solutions to the problems raised. While the organization, sponsorship, and administration of the program have been described in detail elsewhere¹ and are outside the scope of this article, it is necessary to make at least a summary statement concerning them in order to orient the reader to the subsequent discussion.

The National Clerical Ability Testing Program is sponsored by the National Office Management Association and the National Council for Business Education. Its purpose is to appraise the fitness of high school, business school, and college graduates for beginning office positions in the fields of stenography, typing, machine transcription, bookkeeping, calculating

¹*National Clerical Ability Tests*. Bulletin No. 1, November 1939. Joint Committee on Tests. (Cambridge, Mass.: Harvard University.)

National Clerical Ability Tests. Report on 1940 Testing Program. (*ibid.* 1940.)

W. J. E. Crissy, "The Testing Program of the Joint Committee of the National Office Management Association and the Business Educational Council." *Conference of State Testing Leaders* (Proceedings of) October 28, 1939. (Washington, D. C.: American Council on Education).

machine operation, and filing. To assist in such an appraisal, the *National Clerical Ability Tests* are administered annually in centers throughout the country. In addition to tests of skill in the fields referred to, a basic battery of tests is used to measure the prospective employee's competence in English, arithmetic, business information, and general information. Certificates are awarded which are based upon the examinee's proficiency demonstrated on the special skills tests and also upon his general background as measured by the basic battery. Most examinees are candidates for just one certificate and take only one of the skills tests. However, as many as three certificates may be sought by each examinee. Certificates are awarded in each field in which the candidate reaches the required proficiency.

The procedures and problems involved in this program will be discussed under the following heads:

- (1) The Separate Tests—a description of their form, content, etc.
- (2) Statistical Methodology—differential weighting of each test for each group of candidates in a given skills field.
- (3) Certification Procedure.

The Separate Tests

All tests in the basic battery in 1941 are of the objective type and are scored by machine. The tests in English, arithmetic, and business information are printed in a single booklet called the *Fundamentals Test*. This test requires 90 minutes of testing time. The areas of English sampled include spelling, word usage, and the use of the apostrophe in possessives and contractions. An improvement in the English Test would be the inclusion of a section measuring the examinee's knowledge of punctuation (rather than only one aspect of it) and a section testing vocabulary. The Arithmetic Test comprises problems involving the four basic arithmetic operations, and applied problems, such as the handling of discounts and the computation of interest. All items in this test are in five-choice

NATIONAL CLERICAL ABILITY TESTING PROGRAM

form. The choices include four plausible answers and a fifth alternative, "*None of the above.*" The examinee must actually compute the answer to each problem since in certain items all four of the plausible answers are incorrect. The Business Information Test samples the applicant's knowledge of office procedures, postal regulations, technical business terms, and their applications. The General Information Test measures the examinee's knowledge in such areas as world affairs, sports, etiquette, geography, and history. It requires 50 minutes of testing time. This test has a wide range of discriminability and is positively correlated with every test in the battery, yet the correlations indicate some independence of measurement. It has been suggested that this test be replaced by a test of general intelligence, but various personnel officers have reported favorably on its inclusion in the battery. There is some evidence from the use of the test in employment offices to indicate that it has some value in predicting successful adjustment on the job.

The tests in the skills fields are miniature tests, that is, they present in miniature typical business situations in each of the areas included. They are long enough to measure the speed and accuracy with which particular tasks can be done over a significant period of time.

The *Stenography Test* provides for 48 minutes of dictation and 120 minutes of transcription. Fifteen items are dictated including letters and memoranda to be edited. Examinees are furnished printed copies of the letters to which the dictated letters are replies. A relatively even speed of dictation is maintained; the rate of dictation is 90 words per minute. Unusual spelling or punctuation is explained, and requests to repeat sections of the dictation are heeded if made within a specified time. The administration procedures are a departure from the usual school test in stenography but they are in accordance with usual office conditions. In the transcription similar steps have been taken to approximate business practice: erasures are permitted, change of wording is

not penalized when the sense of the item is kept, small point deductions are made for correctible errors, e.g., transposing letters, while severe penalties are imposed for uncorrectible errors such as omissions that would require interlineation. In computing the total score a bonus, proportional to the number of minutes remaining, is given to all candidates who hand in their transcriptions before "time" is called. The chief problem in connection with this test is concerned with scoring. It seems logical that, in terms of office practice, speed does not become advantageous until some acceptable level of accuracy is reached. Under the present plan of scoring, no such level of accuracy has been specified and hence some examinees obtain high scores due to their typing speed while their accuracy is at a level of doubtful acceptance in beginning office work.

The *Typing Test* permits a maximum of 120 minutes of working time. Form letters, reply cards, etc., are furnished the examinee, and he is given several typing jobs which involve the use of the materials furnished. This test approximates office conditions, as does the *Stenography Test*, and it provides a more comprehensive measure of typing ability than can be obtained from the usual kind of test in this field. However, the scoring problem is the same here as in the *Stenography Test*. No provision is made for a minimum accuracy score below which a bonus for time saved may not be added when the total score is computed.

The *Machine Transcription Test* involves the transcription of seven items from either an Ediphone or Dictaphone cylinder. A maximum of 60 minutes is allowed for the transcription. Scoring procedures are in line with business practice except, again, for the inadequate method of handling the speed aspect of the score.

The *Bookkeeping Test* requires the examinee to work to completion specified operations on excerpts from a set of books. The format of this test has been approved and recommended by experts in bookkeeping and accounting procedures. There is evidence to indicate that it measures bookkeeping

NATIONAL CLERICAL ABILITY TESTING PROGRAM

more adequately than do tests involving indirect evidences of ability in this field.

The *Machine Calculation Test* measures the examinee's ability to carry out the four basic arithmetic operations on a key-driven calculating machine. The addition problems cover columnar summing and cross-footing. The multiplication section ranges from simple multiplication to multiplying to obtain a sum of the products. The subtraction problems extend from the very simplest single subtractions through problems requiring first alternate columnar summing and then subtractions to obtain balances. The division section includes questions in direct division and also in obtaining reciprocals to be used as multipliers. The entire test requires 120 minutes of working time.

The *Filing Test* samples the examinee's ability to file various materials furnished and also his knowledge of acceptable filing procedures in the solution of problems that frequently occur in office practice. Alternative sections are included in the last part of the test to cover different filing systems. The test requires 120 minutes of working time.

Statistical Methodology

In order to obtain an over-all appraisal of the examinee's ability, scores on the basic battery and the skills test are combined into a "best-weighted" composite. Since the competencies measured by the various tests in the basic battery are of different importance in each of the six skills fields, there exist six different weighting applications, one for each group of examinees taking each of the skills tests.

Four components make up the weight accorded to each test (both basic battery tests and skills test) within each of the six fields:

- (1) The variability or dispersion of the scores made by the group of examinees.
- (2) A function of the test's reliability for the group; the quasi-regression weight of the test.
- (3) The estimated importance of the test.
- (4) The independence or uniqueness of the test.

The method of combining these components may be symbolized thus:

$$W_{ij} = \beta'_{ij} \left(\frac{I_{ij}}{\sigma_{ij}} + U_{ij} \right)$$

where

W_{ij} = weight accorded test i within a particular skills battery j (a skills battery includes a skills test and the basic battery);

β_{ij} = quasi-regression weight of the test i in battery j ;

I_{ij} = estimated importance of the competency measured by test i in the battery j ;

U_{ij} = uniqueness of test i in battery j ;

σ_{ij} = standard deviation of scores on the test i of persons selecting the skills test designated for battery j .

Standard deviations are computed within skills groups for each test and are used in the weighting procedures as indicated in the formula presented above. This furnishes the first component of the weight for each test in each battery.

No criteria are now available² against which to obtain regression weights on the tests. However, to account for the second component, quasi-regression weights are computed using Kelley's formula³: $\beta' = \frac{\sqrt{r_{ii}}}{1 - r_{ii}}$, where r_{ii} is the reliability coefficient of the test.

The third component of each test weight is determined by having a committee of experts in each skills field independently judge the importance of the competency measured by each of the basic tests (English, arithmetic, business information, and general information) relative to a basic weight of 10 accorded the skills test in that field. For example, in stenography, the importance weights are:

English 3

Arithmetic 1

²Mr Robert Blanton has an extensive validation study in progress involving 1939 and 1940 examinees.

³For the derivation and rationale of this formula see: T. L. Kelley, *Interpretation of Educational Measurements* (Yonkers: World Book Company, 1927), pp 212-213.

NATIONAL CLERICAL ABILITY TESTING PROGRAM

Business Information 2

General Information 3

Stenography 10

Uniqueness, the fourth component, is measured by using the median alienation coefficient for each test within a particular battery. This involves obtaining six intercorrelation matrices, each matrix including a particular skills test and the basic tests (5 x 5 matrix). The median correlation coefficient in each row is then used to obtain the alienation coefficient indicated above.

In order to "equalize" weights for *importance* and *uniqueness* so that the sums of the two sets of weights will be equal before the several components of the weight, W'_{ij} , are combined, the alienation coefficients (uniqueness components) are each multiplied by a constant equal to the columnar sum of the five *importance* weights divided by the columnar sum of the five alienation coefficients. This is done in the case of each of the six skills batteries.

To illustrate the weighting procedure, the computational steps in the case of the 1941 Stenography battery are indicated in Table 1.

TABLE 1
STENOGRAPHY

A	B	C	C ¹	D	E	F	G	H	I	J
Stenography	46.55	.90	9.487	.98	10	3.995	132.771	2.852	3.08	
Bus. Inf.	11.48	.66	.59	1.873	.90	2	3.669	10.618	.925	1.00
English	7.66	.84	.70	2.789	.93	3	3.792	18.943	2.473	2.67
Bus. Arith.	4.29	.75	.73	3.164	.93	1	3.792	15.162	3.534	3.82
Gen. Inf.	22.31	.84	.82	5.031	.92	3	3.751	33.964	1.522	1.65

$\Sigma E = 4.66$ $\Sigma F = 19$

Column Data Presented

- A Tests included in Stenography battery.
- B Standard deviation for each test within the Stenography group (σ_{ij}).
- C Reliability coefficients computed by "split-half" method based upon sample.

- C¹ Adjusted reliability coefficients (corrected to range of Stenography group) by formula:

$$r_{11} = 1 - \frac{\Sigma^2}{\sigma^2} (1 - R_{11})$$

- D Quasi-regression weights (β'_{iu}).
 E Uniqueness weights (median k 's).
 F Judges' weights of importance (I_{iu}).
 G Entries in column E adjusted relative to entries in column F by the formula:

$$G = \frac{\Sigma F}{\Sigma E} E \quad (U_{iu})$$

- H $D(F + G) = \beta'_{iu} (I_{iu} + U_{iu})$

$$I \quad \frac{H}{B} = \frac{\beta'_{iu} (I_{iu} + U_{iu})}{\sigma_{iu}} = H'_{iu}$$

- J Entries in I, each divided by smallest entry in I (Column J contains the weights which are finally used. This makes subsequent computation easier by making the smallest H'_{iu} equal to unity.)

When the weight for each test has been obtained by the foregoing procedures, a composite score for each examinee is obtained by multiplying each score by the appropriate weight and summing his weighted scores.

Certification Procedure

The certification of an examinee in a particular skills field depends upon two factors (1) having a skills test score equal to or in excess of the critical or "passing" score set for that particular skills test; and, in addition, (2) having a composite score equal to or in excess of the critical or "passing" score set for that particular composite.

The critical score for each skills test is established by a committee of experts in that field (usually the same committee that estimates the *importance* weights). The criterion used by each committee is "minimum acceptable performance in a beginning office job." The procedure used is to have each committee member inspect and judge as acceptable or unacceptable a sample paper from each two per cent segment of the

NATIONAL CLERICAL ABILITY TESTING PROGRAM

distribution beginning at the twentieth percentile point and extending to the eightieth percentile point. After independent judgments are completed, the combined judgment of the committee is used to determine the critical score on the particular skills test.

To determine the critical composite score, regression technique is employed; the desired critical composite score is predicted from the previously determined critical skills score through regression of composite on skill.

Obviously the correlation between each skills test and the corresponding composite is high, since the skills test is the chief component of the composite. An improvement in this procedure would be to exclude the skills test from the composite and thus use the composite as a "background index." Then if the critical composite score were obtained by prediction from the critical skills test score by means of a regression equation involving composite and skill, the only hypothesis involved would be that the minimum "background" score should be about equivalent to the minimum skills score.

Summary

In this paper have been discussed the measurement procedures and problems connected with the National Clerical Ability Testing Program. The treatment of problems has been limited to those which are peculiar to this particular program. The procedures, however, have been covered in detail because they have general application to other types of testing projects.

The weighting and certification procedures described in this paper should obviously be completely revised as soon as "outside criteria" are available against which to weight the tests. Probably the best procedure to use when these criteria are available is canonical correlation technique modified to include *importance* weightings of both the separate criteria and the separate tests.

So long as such outside criteria are not available, the pro-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

cedures used at present should be modified either in accordance with the suggestions made in this paper or in some other manner if the program is to render increased service to prospective clerical employees and to the employers of such persons.

INTROVERSION-EXTROVERSION AS A FACTOR IN TEACHER-TRAINING

CATHARINE EVANS

Indiana University

and

C. GILBERT WRENN

University of Minnesota

Introduction

ONE OF THE SERIOUS problems facing teacher-training institutions today is the selection of student personnel. Many teachers are unsatisfactory either because of inadequate training or unfortunate personality characteristics. Teacher-training institutions must select students who can benefit most from the improved training programs now provided. This study is intended to throw some light on the relationship of personality traits to student success in teacher-training programs. More specifically, the purpose of this study has been to determine the relationship of Introversion-Extroversion¹ to the scholastic achievement and student teaching success of education students. An *I-E Inventory* was administered to 396 seniors in the College of Education at the University of Minnesota. This inventory will be described briefly in order that the results of the study may be understood clearly. A more complete discussion of the construction of the inventory is available in a recent article in the *Journal of Psychology* (1).

This inventory was constructed to measure three types of

¹Throughout the remainder of the article, Introversion-Extroversion will be designated as I-E.

I-E, Thinking, Social, and Emotional, which were isolated by Guilford (2) in his factor analysis of I-E. Original items were developed and stated in the form of questions concerning the behavior and reactions of the student. The questions were formulated in such a manner that the student could indicate how frequently he or she behaved in that way. Typical questions were, "Do you question statements and ideas expressed by your professors?" "Do you enjoy eating meals alone?" and "Do you avoid exaggeration in your statements?"

The construction and choice of items for the three tests in the inventory were guided by the following definitions which contrast the extremes for each type of I-E:

The thinking introvert likes reflective thought, particularly that of a more abstract nature. His thinking tends to be less dominated by objective conditions and generally accepted ideas than thinking of the extrovert. The thinking extrovert, however, shows a liking for overt action, and his ideas tend to be ideas of overt action.

The social introvert withdraws from social contacts and responsibilities and displays little interest in people. In contrast, the social extrovert seeks social contacts and depends upon them for his satisfaction.

The emotional introvert tends to repress and inhibit outward expression of his emotions and feelings. On the other hand, the emotional extrovert readily expresses his emotions and feelings outwardly. He shows a greater tendency to make the expected response to simple, direct emotional appeals than the introvert.

In constructing this inventory an effort was made to develop relatively independent measures for these three types of I-E by a technique of item analysis. The intercorrelation coefficients for the 396 College of Education seniors were: Thinking and Social I-E tests, $-.25$; Thinking and Emotional I-E test, $+.17$, and Social and Emotional I-E tests, $-.23$. These low coefficients indicate that the three tests are relatively independent. The inventory also seems sufficiently reliable for individual prediction since each of the tests has a reliability coefficient with groups of education students of $.88$ or above,

for either the retest or split-half technique or for both techniques.

Indirect evidence concerning the validity of each test is available, in terms of the ability of the test to differentiate known groups of college students which on an *a priori* basis seemed to be extreme in the type of I-E involved. For example, the Thinking I-E test significantly differentiated major groups in the College of Education; the majors in physical education, home economics, commercial education, and child welfare were extreme in the direction of Thinking Extroversion, while the majors in English, art, mathematics, social studies, and languages were extreme in the direction of Thinking Introversion. The Social I-E test significantly differentiated groups of students varying in the degree of participation in campus activities; the members of academic sororities and fraternities and the students active in campus organizations were found to be more socially extroverted than the non-affiliates and non-participants. Likewise, the Emotional I-E test significantly differentiated sex and age groups; women were more emotionally extroverted than men, and the younger student groups were more emotionally extroverted than older groups. Each test did differentiate known groups of students which logically were expected to be extreme in that type of I-E.

Scholastic Achievement of Student Groups Varying in Thinking I-E

The relationships of the scores on the three tests in the *I-E Inventory* to measures of scholastic and student teaching success have been explored in this study of College of Education seniors.

The relationship of Thinking I-E to scholastic achievement honor point ratios was explored for the seniors in the College of Education who had taken a scholastic aptitude test, the *Miller Analogies Test*, during the junior or senior year. The *Miller Analogies Test*—Form G, consists of 100 analogies which research with college students indicated were discriminating items. Data reported by Dugan (3) indicate that

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

this test is highly reliable and valid as a measure of scholastic aptitude. The correlation between the scores on the Thinking I-E and the *Miller Analogies Tests* varied from .15 to .26 with groups of 112 to 260 education students, indicating only a small positive relationship.

Significant results were obtained in the study of the scholastic achievement of 148 native students, i.e., students who had taken all of their college work at the University of Minnesota. These students were divided into four groups according to the varying size of their scores on the Thinking I-E test, and the mean honor point ratios for each group were computed for the total course work, for major courses, and for education courses. There was in general a progressive increase in the mean major and education honor point ratios with an increase in introversion in this group. Those students with Thinking I-E scores in the quarter extreme in the direction of introversion had significantly higher mean honor point ratios than those with scores in the quarter extreme in the direction of extroversion (see Table 1).

TABLE 1
MEAN HONOR POINT RATIOS IN TOTAL, MAJOR, AND EDUCATION COURSES
FOR FOUR GROUPS OF NATIVE STUDENTS VARYING IN
DEGREE OF THINKING I-E

Thinking I-E Groups	Mean Honor Point Ratio		
	Total	Major	Education
Upper Quarter (Extroversion)	1.3711	1.6743	1.5557
Second Quarter	1.5211	1.7878	1.8050
Third Quarter	1.6776	1.9051	1.8849
Lowest Quarter (Introversion)	1.6138	1.9292	1.8926

TABLE 2
DIFFERENCE AND THE SIGNIFICANCE OF THE DIFFERENCE IN MEAN
HONOR POINT RATIOS FOR THE EXTREME THINKING
I-E QUARTILE GROUPS OF NATIVE STUDENTS

Variable	Difference		Probability of <i>t</i>
	in means	<i>t</i>	
Total Honor Point Ratio2427	2.9241	.01
Major Honor Point Ratio2549	2.7117	.01
Education Honor Point Ratio3369	2.9553	.01

INTROVERSION-EXTROVERSION IN TEACHER-TRAINING

The analysis of variance technique (4) was applied to the total, major, and education honor point ratios for the four Thinking I-E groups in order to test the significance of the difference in means. The variance among the mean honor point ratios of the Thinking I-E groups was significantly larger for each of the three analyses of variance than the variance within the groups. The ratio of the variances for each type of honor point ratio satisfied at least the five per cent level of significance. The results of the analysis of variance for each type of honor point ratio refuted the hypothesis that there was no difference in the scholastic achievement of the four Thinking I-E groups. The groups were heterogeneous in scholastic achievement (see Table 2).

The analysis of variance technique with the two criteria of classification, Thinking I-E and *Miller Analogies* scores, was also employed with the total honor point ratios in order to determine whether or not the variance among the mean honor point ratios of the four Thinking I-E groups would be significant when the groups were subdivided according to *Analogies* ability.

The variance between the Thinking I-E groups was not significantly greater than the variance within the Thinking-*Analogies* subclasses. When the variance within the *Analogies* quartile groups was considered, there was insufficient evidence to determine any difference in the scholastic achievement of the four Thinking I-E quartile groups.

From the analysis of variance data the mean honor point ratios of the following four groups of native students varying both in degree of Thinking I-E *Analogies* ability were compared:

- (1) Students below the median in the direction of Thinking Introversion and above the median in the ability measured by the *Analogies* Test.
- (2) Students above the median in the direction of Thinking Extroversion and above the median in *Analogies* ability.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

- (3) Students below the median in the direction of Thinking Introversion and below the median in Analogies ability.
- (4) Students above the median in the direction of Thinking Extroversion and below the median in Analogies ability.

A steady decrease in mean total honor point ratio from the first to the fourth groups can be noted in Table 3.

TABLE 3

MEAN TOTAL HONOR POINT RATIOS FOR FOUR GROUPS OF NATIVE STUDENTS VARYING IN THE DEGREE OF THINKING I-E AND OF ILL. ABILITY MEASURED BY THE MILLER ANALOGIES TEST

Number of Group	Type of Groups	Number	Mean Total Honor Point Ratio
1	<i>Below</i> the median in the direction of Thinking Introversion and <i>above</i> the median in Analogies ability.....	46	1.7565
2	<i>Above</i> the median in the direction of Thinking Extroversion and <i>above</i> the median in Analogies ability.....	28	1.5218
3	<i>Below</i> the median in the direction of Thinking Introversion and <i>below</i> the median in Analogies ability.....	28	1.4636
4	<i>Above</i> the median in the direction of Thinking Extroversion and <i>below</i> the median in Analogies ability.....	46	1.4000

The first group had a significantly higher mean honor point ratio than the second group ($t = 2.24$). However, this second group did not have a significantly higher mean honor point ratio than the third ($t = .58$). These data seem to indicate that high Analogies ability combined with a tendency toward Thinking Introversion is more ideal from the standpoint of scholastic achievement than either high Analogies ability combined with a tendency toward Thinking Extroversion or low Analogies ability combined with a tendency toward Thinking Introversion.

The question of the relation of Thinking Introversion and high Analogies ability to scholastic achievement was attacked

INTROVERSION-EXTROVERSION IN TEACHER TRAINING

from another angle. The 24 native students who had honor point ratios above 2.00, and the 26 who had honor point ratios below 1.50 were compared in mean scores on the Thinking and Analogies Tests (Table 4). The students with the higher honor point ratios had significantly higher Thinking Introversion scores and Analogies scores than those students with lower honor point ratios. The values of "*t*" were 2.51 and 3.58, respectively, and they satisfied at least the five per cent level of significance.

TABLE 4

MEAN THINKING AND ANALOGIES SCORES OF NATIVE STUDENTS WITH HONOR POINT RATIOS ABOVE 2.00 AND OF NATIVE STUDENTS WITH HONOR POINT RATIOS BELOW 1.50

Type of Group	No.	Mean Thinking I-E Score [†]	Mean Analogies Score
Natives with Honor Point Ratios above 2.00	24	6.8333	60.0833
Natives with Honor Point Ratios below 1.50	26	23.4615	45.6538

[†]The smaller the score, the greater the tendency toward introversion.

The evidence for the relationship of Thinking Introversion to scholastic achievement for native students seems weakened by the non-significant results obtained in the analysis of variance by the double criteria of classification. However, the study of group differences points to the desirability of the combination of a tendency to Thinking Introversion with high Analogies ability. Indeed, the results of the analysis of variance by the double criteria of Thinking I-E and Analogies ability can be interpreted as strengthening the evidence for the greater desirability of the combination of Thinking Introversion with high Analogies ability in contrast to the desirability of a tendency to Thinking Introversion alone regardless of Analogies ability.

When transfer students, i.e., those students transferring to the University of Minnesota with advanced standing from

other institutions, were studied, the results were not significant. Although the mean honor point ratios of the transfer students in the quartile group extreme in the direction of Thinking Introversion were larger than the means of those in the quarter extreme in the direction of extroversion, the differences were not significant. Likewise, the results of the analysis of variance were not significant. However, significant differences were found between samples of native and transfer students in the distribution by major fields, in the mean scores on the Analogies test and in the means of the major and education honor point ratios. The explanation of the contrasting results obtained with native and transfer students must lie in these differences. It may be safely concluded however, that there is a relationship between Thinking Introversion and scholastic success for the native student in the College of Education.

*The Student Teaching Success of Groups of
Students Varying in I-E*

The relationships of scores on the three I-E tests to student teaching success at the University of Minnesota were also explored. Two measures of student teaching success were employed in this study. In the first place, the rank order lists of the student teachers in 16 major fields were obtained. These rank order lists were made out by the combined group of critic teachers for a major field. Second, the marks in student teaching for the three quarters were computed as an honor point ratio for 242 seniors.

The coefficients of correlation between the student teaching ranks and the ranks of the scores on the three I-E tests were computed for the majors in eight of the 16 teaching fields. These were the majors which numbered 14 or more cases. The Spearman rank difference formula was employed in the calculation of these rank correlation coefficients.

These coefficients as given in Table 5 varied from .00 to .43. They were based on such small samplings that it was improbable that any of the coefficients were significant. All of the coefficients of correlation between Social I-E and student

INTROVERSION-EXTROVERSION IN TEACHER-TRAINING

TABLE 5

RANK COEFFICIENTS OF CORRELATION BETWEEN STUDENT TEACHING RANKS AND THE THREE I-E TESTS FOR EIGHT MAJOR FIELDS

Major Field	N	Thinking I-E Test	Social I-E Test	Emotional I-E Test
1 English	50	-.23	+.19	+.24
2 Social Science	49	-.06	+.17	+.23
3 Child Welfare	38	-.20	+.12	-.03
4 Science	19	-.04	+.38	-.38
5 Commercial	19	-.18	+.43	-.12
6 Art	17	-.37	+.19	+.16
7 Girl's Physical Education	15	-.37	+.35	-.41
8 Mathematics	14	.00	+.25	+.18

teaching rank, however, were positive. This consistent positive relationship of Social Extroversion and student teaching rank for the eight major fields does seem to indicate the type of relationship existing between these two factors. Six of the eight coefficients between student teaching ranks and the scores on the *Thinking I-E Test* were negative. Thus a tendency was indicated for Thinking Introversion to be related to student teaching success. The relationship of Emotional I-E to student teaching rank was not consistent in the eight major fields. Four coefficients were positive, and four were negative.

The relationship of student teaching success to the I-E scores was also determined for 55 students whose student teaching ranks were in the upper and lower quarters on the rank order lists of the same eight major fields. This analysis indicated that the more successful student teachers tended toward more thinking introversion, social extroversion, and emotional extroversion in central tendency than the less successful student teachers. However, only one of these differences in mean I-E scores was significant. The students in the upper fourth on student teaching rank order lists were significantly more socially extroverted than the students in the lower fourth; the value of "*t*," 2.58, satisfied the five per cent level of significance. There were, on the other hand, more than five chances out of one hundred that the differences in the

mean Emotional and Thinking scores could have occurred from chance errors of sampling.

Since there was a significant difference in the mean scores on the Social I-E test for the groups in the upper and lower fourths on the student teaching rank order lists, the analysis of variance technique was employed to study these differences. The mean scores on the Social I-E test of the four groups formed on the basis of ranks in student teaching are given in Table 6. There was a progressive increase in mean scores on

TABLE 6

MEAN SCORES ON THE SOCIAL I-E TEST OF THE FOUR GROUPS FORMED ON THE BASIS OF RANK IN STUDENT TEACHING

Groups Varying in Student Teaching Rank	No.	Mean Social Score*
Upper Quarter	55	15.9273
Second Quarter	57	15.2632
Third Quarter	54	7.7963
Lower Quarter	55	4.0727

*The larger the score, the greater the degree of extroversion.

the Social I-E test for the four groups with the increase in student teaching rank. In other words, the tendency to Social Extroversion increased as student teaching rank became higher. The simple analysis of variance was employed to test the null hypothesis that there was no difference in the four groups from which the mean Social I-E scores were obtained. The variance among the four groups varying in practice teaching ranks was significantly greater than the variance within the groups, indicating that the four groups were not homogeneous in Social I-E. These data indicated that Social Extroversion was related to student teaching success as measured by the ranks of critic teachers. The more successful the student teacher, the greater was the tendency to Social Extroversion.

The marks in student teaching for three quarters of work were also employed as a criterion for the choice of two groups of students of extreme degrees of success in teaching. A

INTROVERSION-EXTROVERSION IN TEACHER-TRAINING

majority of the seniors had student teaching honor point ratios between 2.00 and 2.50. There were 68 students with ratios above 2.50 and 67 students with ratios below 2.00.

The mean scores on the I-E tests of these two groups were compared. The same differences were found for these two groups as for the two groups chosen by the criterion of student teaching ranks with the exception of the scores on the Thinking I-E test. The mean scores indicated that the more successful student teachers as judged by their marks were more extroverted in Thinking, Social, and Emotional I-E than the less successful teachers. However, the values of "*t*" were so small that none of the three differences in means was significant. The rank order lists seemed to yield a more differentiating measure of student teaching success than marks in student teaching. In fact, students with a B average (2.00 ratio) in student teaching varied in the rank given by critic teachers from the lowest to the upper quarter.

The results of the use of the two criteria for teaching success indicate that the more successful student teacher is on the average more extroverted, socially and emotionally, than the less successful student teacher. No conclusion can be made in relation to the Thinking I-E test because of the conflicting results.

Summary

The results of this study indicate that for "native" seniors in the College of Education, Thinking Introversion is related to high scholastic achievement and that Social and Emotional Extroversion are related to student teaching success. These results provide evidence to substantiate a common "hunch" that good teachers are not only good students but also must possess certain social and emotional characteristics. The results indicate that a combination of high mental ability and Thinking Introversion is desirable for scholastic success. In addition, Social Extroversion is also necessary for high rank in student teaching. The extent to which the I-E Inventory²

²This inventory will be published soon by Science Research Associates.

can be used to predict scholastic success in the College of Education or student teaching success has not yet been determined. The I-E Inventory is being used, however, in a current study in this same college that will follow a class of juniors through a four-year period. By studying their success in student teaching and on the job, it will be possible to indicate the predictive values of the inventory and other instruments not only for the training period but also for job adjustment.

REFERENCES

1. C. Evans and T. R. McConnell. "A New Measure of Introversion-Extroversion," *Journal of Psychology*, XII (1941), 111-124.
2. J. P. Guilford and R. B. Guilford. "Personality Factors, S, E, and M, and Their Measurement." *Journal of Psychology*, II (1936), 109-127.
3. Willis E. Dugan. "A Study of the Miller Analogies Test with Graduate Students in the College of Education." Unpublished Master's Thesis. University of Minnesota, 1939.
4. George W. Snedecor. *Statistical Methods*. Ames, Iowa: Collegiate Press Inc., 1938, 387 pp.

AN INVESTIGATION OF THE POSSIBILITIES OF MEASURING PERSONALITY TRAITS WITH THE STRONG VOCATIONAL INTEREST BLANK¹

LYLE TUSSING

Wilson Junior College

THE *Strong Vocational Interest Blank* is a widely used instrument for determining vocational interest patterns in educational and vocational guidance. This study was contemplated because it was believed that there was a possibility of weighting items on the *Interest Blank* in such a way as to obtain, with this single test, certain personality measures as well as the vocational interest scores now available. With this object in mind, the present study analyzes the relationship between responses on the *Strong Vocational Interest Blank* and scores on certain personality tests to determine how well the traits measured by these tests can be measured by the Strong Blank.

The idea of evaluating other factors than vocational interest with the *Strong Vocational Interest Blank* is not new. Strong has used his Blank to measure masculinity and femininity (20) and also interest maturity (18, 19). Young and Estabrooks (21) studied the relation of personality and interest tests to scholastic success. They found that the *Strong Vocational Interest Blank* showed evidence of being the most significant predictive measure after they had made an item analysis of several tests.

¹This study is a portion of a thesis submitted to the Faculty of Purdue University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, June, 1941

The purpose of the present study was to construct scoring keys for the *Strong Vocational Interest Blank* to measure certain personality traits, such keys being: (1) suitable for use in group testing and available as another key for the *Strong Vocational Interest Blank*, (2) so constructed that scoring is both rapid and free from the personal error. Validation of the scales was obtained by correlating the scores resulting from the new Strong scoring keys with scores for the corresponding Allport-Vernon, Bell, Bernreuter, and the American Council on Education tests.

The matter of falsification of responses to test items is a problem that is worthy of consideration. With most personality tests, it is a very easy matter for the person being tested to falsify his responses if he wishes. (16) It has also been shown that the *Strong Vocational Interest Blank* can be changed to falsify interest in a vocation (13). However, in most cases where the individual is interested in obtaining guidance, he will not deliberately falsify his responses. Also it seems that because of the number and brevity of the items in the *Strong Vocational Interest Blank*, falsification of responses by the subject possibly would be more difficult than in the conventional personality inventory of the direct question type.

It was not deemed necessary in the course of this investigation to examine the nature of personality traits, to define them, nor to investigate all the possible traits. In this study, the validity and reliability of the tests used was accepted, and no attempt was made to prove or disprove whether the tests selected were valid measures of the traits which they purported to measure. (1-10 and 15)

The exact number and the names of existent personality traits have not been agreed upon Allport (4) states that "The Eugenics Record Office has issued a *Trait Book* containing a list of approximately 3,000 characteristics that might conceivably be hereditary according to the principle of unitary characteristics." Further he cites McDougall as listing five elements of personality; Beck, four; and Boven, three. While

it is unlikely that authorities will agree as to the units of personality and their exact number, nevertheless, measures of "sociability", "confidence in one's self", "home adjustment", "health adjustment", and "emotional adjustment", as well as intelligence, are quite widely measured factors of personality, and consequently these measures have been selected for investigation in the present study.

Procedure

In this study the problem was to determine how the items on the *Strong Vocational Interest Blank* were related to the scores made on the Allport-Vernon, Bell, and Bernreuter personality tests, and on intelligence tests, and also to find whether the Strong Blank could be used as a means of measuring the same personality traits as the above-mentioned tests if the items were weighted. In order to obtain weightings, it was necessary to determine how the groups scoring at the extremes of the measures (high, low) varied in their responses to items on the *Strong Vocational Interest Blank*.

A sample group of 300 men was used for establishing the weights. This group was one originally studied by Dr. E. Lowell Kelly (14) in an investigation of factors contributing to marital happiness.² Consequently, it was a group in which several selective factors were operating. For example, each man was about to be married. He was willing to cooperate in a study in which such factors as intelligent curiosity and cooperativeness play an important role. His general intelligence and educational background were perhaps higher than the average. Most of this group had attended college, and the average age of the group was 26.66 with a standard deviation of 3.47. However, even though these selective factors were operative, the group nevertheless represented a relatively heterogeneous group with respect to the variables under analysis and therefore was satisfactory for the present investigation.

²Data for these studies were collected with the aid of a series of grants from the Committee for Research on Problems of Sex of the National Research Council

The 300 subjects were given a battery of tests including the *Strong Vocational Interest Blank*, the Allport-Vernon *Study of Values*, the Bernreuter *Personality Inventory*, the Bell *Adjustment Inventory*, and the Otis S. I. *Test of Mental Ability*. The present study utilizes their responses to these tests as basic data.

Scores of the 300 men on the nine scales (four scales of the Bell *Adjustment Inventory*; two of the Bernreuter *Personality Inventory*, F1-C and F2-S; the Otis S. I. *Test of Mental Ability*; two scales, theoretical and economic, of the Allport-Vernon *Study of Values*) were coded and punched on cards. This set of cards was then sorted on each of the nine scales to identify the subjects scoring in the highest 25 per cent and the lowest 25 per cent of the scores on each scale. In many instances, due to coding, it was impossible to select exactly 25 per cent (75 cases), although this number was desired in each of the two groups.

After these two extreme scoring groups were obtained, the next step was that of comparing the responses of these groups for each item on the Strong Blank. This was done by punching the responses of the 300 subjects to the 420 items of the 1927 form of the *Strong Vocational Interest Blank* on Powers cards. This punching consisted of recording the responses for each individual on each item as to "like", "indifferent", or "dislike" for the particular item. In some instances the subject failed to respond to an item, but, since inspection showed these omissions to be fairly equally distributed in the high and low groups, it was considered advisable not to weight such omissions.

After determining which individuals constituted the high and low groups of a particular test, the cards containing the Strong responses of the individuals composing each group were hand sorted by serial number of the subject. Each resulting "high" group of cards was then sorted by item and the number of persons answering "like", "indifferent", and "dislike" for each of the 420 items in the high group was recorded. This

same procedure was followed in the low group. The percentages were then obtained for each of the 420 items for "like", "indifferent", and "dislike" for both the high and low groups (2,620 percentages for each scale). This procedure was followed for the nine scales.

After the percentages had been computed for the high and low groups on "like", "indifferent", and "dislike", the difference in the two percentages for "like", "indifferent", and "dislike" for each item was compared and the differences in percentage were given a weighting. These weightings ranged from +4 to -4 according to the formula
$$W = 100 \frac{\Delta}{1 + 4j \Delta^2}$$
 reported by Strong (12, 17) to be the most satisfactory scheme he had found for assigning weights to his interest scales.

These new weights were then transferred to the 1938 Strong Blank. Since the 1938 form included only 400 of the original 420 items, some of the responses had to be omitted in making this transfer. However, the number of changes was small and probably did not appreciably affect the reliability or validity of the new scales. The weights as they now appeared on the 1938 form of the *Strong Vocational Interest Blank* constituted a new scoring key. These weights were then transferred to International Business Machines test-scoring keys.

A new sample of 103 male college freshmen and sophomores was used as a validating group. These students took the battery of tests consisting of the *Strong Vocational Interest Blank*, the Allport-Vernon *Study of Values*, the Bernreuter *Personality Inventory*, the Bell *Adjustment Inventory*, and the American Council on Education *Psychological Examination* voluntarily. Since considerable time was consumed by each individual participating, not all of the men finished all of the tests in the battery. However, as many subjects as possible were utilized in the validation of each new scoring key.

The responses of the validating group on the *Strong Vocational Interest Blank* were then scored with each of the various

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

new keys (i.e., the key for "economic",¹ "theoretical", "intelligence", etc.). By scoring the odd and even items separately, it was possible to compute the reliability for each new scale. The scores made on the "odds" were correlated with those made on the "evens", and the reliability for the whole test was estimated by means of the Spearman-Brown formula.

The validity was obtained by correlating the total (odd plus even) scores on the *Strong Vocational Interest Blank*; for example, the total scores on "theoretical" with the theoretical scores obtained on the Allport-Vernon *Study of Values*.

Results

Table 1 indicates the number of weighted items on the *Strong Vocational Interest Blank*. It also shows the results of the validating group for the Bernreuter (F1-C, F2-S Scales), Allport-Vernon (Theoretical, Economic Scales), Bell (Home, Health, Social and Emotional Adjustment Scales), and Intelligence as measured by the American Council on Education *Psychological Examination*.

TABLE 1
RESULTS OF THE NEW STRONG BLANK SCALES

	"F1-C"	"F2-S"	"Theor."	"Econ."	"Home Adj."	"Health Adj."	"Social Adj."	"Emot Adj."	"Intel"
Mean	-17.4	19.5	9.3	4.2	-12.8	-5.0	+8.5	-6.7	23.7
S. D.	58.2	36.1	64.3	64.9	38.2	22.2	54.6	25.3	53.4
Reliability	.86	.77	.82	.80	.83	.70	.87	.77	.80
Validity	.48	.52	.48	.56	.21	.34	.50	.07	.45
No. Wgtd Items	347	327	336	325	319	246	339	289	343
No Items Wgtd. 2 or more	130	70	143	118	54	24	108	34	129

¹To distinguish scores on the new from those on the original scales, each of the newly derived scales is designated by using the same name or symbol as for the original test, but is enclosed in quotation marks.

It is interesting to note the items on the Strong Blank which compose these scales. Some of the better (more heavily weighted) items on the Strong "F1-C" Scale (The Beinreuter F1-C Scale purports to differentiate the self-conscious individual from the self-confident individual) which appeared to differentiate the *self-conscious* individual from the *self-confident* person show that the former dislikes the occupation of "aviator". He dislikes the school subjects "physical training" and "public speaking." In the field of amusements, he likes "picnics", but dislikes "rough house invitations." With reference to activities, he dislikes "adjusting a carburetor", "repairing electric wiring", "making a speech", "organizing a play", "opening a conversation with a stranger", "acting as a yell-leader", and "expressing judgments publicly regardless of criticism." In the section dealing with peculiarities of people, he dislikes "people who assume leadership", but likes "talkative people." He would rather listen to a story than tell a story. In rating his present abilities and characteristics, he does not usually start the activities of a group, work steadily, or liven the group on a dull day. He is not quite sure of himself, and he is doubtful as to his ability to accept criticism without getting sore and to distinguish between more or less important matters. He does not discuss his ideals with others, and his feelings are easily hurt. He loses his temper at times and worries considerably about mistakes.

Opposed to this, the self-confident individual likes "meeting and directing people", but he dislikes "talkative people." He is indifferent as to whether he would rather tell a story or listen to a story, and he likewise has no preference in the matter of a few intimate friends or many acquaintances. He answers in the affirmative to "am quite sure of myself", to "accept criticism without getting sore", and "discuss my ideals with others." He does not get rattled easily, and his feelings are not easily hurt. The person who is self-confident "enters into the situation and enthusiastically carries out the program", and he does not worry.

The Bernreuter F2-S Scale is a measure of sociability. Some of the better (i.e., more heavily weighted) items on the Strong "F2-S" which appeared to differentiate the *non-social* individual from the *social* individual show that the non-social individual likes to "deal with things" rather than people. He would like to be a magazine writer, and he prefers "amusements alone or with two or three others" rather than "amusement where there is a crowd." He would rather spend nights at home than away from home, and he enjoys reading a book rather than going to the movies. The non-social individual reports that he has a "few intimate friends" rather than "many acquaintances." He says he does not "win friends easily", nor does he "usually liven up the group on a dull day." He "can write a concise, well-organized report." Such an individual "practically never tells jokes" and his "feelings are easily hurt."

The social individual is one who "tells jokes well." He is also a person who is indifferent in his choice between a "great variety of work" or a "similarity of work."

The Allport-Vernon Theoretical Scale was designed to segregate individuals whose theoretical values in life are highest. On the new Strong "Theoretical" Scale, there are a number of more heavily weighted items which appear to differentiate the *theoretical* individual from the *non-theoretical* type. For example, in indicating the occupations he would prefer, the former names as "likes" the following: "astronomer", "author of a technical book", "chemist", "inventor", "laboratory technician", "marine engineer", "scientific research worker", "statistician", and "watchmaker." The theoretical person is fairly consistent in preferring such school subjects as "calculus", "geology", "philosophy", "physics", "physiology", and "zoology." The theoretical type of individual enjoys "museums" and the "solving of mechanical puzzles", also "doing research work." In the section of the Strong Blank devoted to the order of preference of activities, he indicates "like" for the development of the theory of operation of a

new machine and for the role of chairman of an educational committee. He would have liked to have been a "Luther Burbank" or a "Thomas Edison." He likes "technical responsibility (head of a department of 25 people engaged in technical, research work)" as opposed to "supervisory responsibility (head of a department of 300 people engaged in typical business operation)", and he would prefer "mental activity" to "physical activity." In rating his own abilities and characteristics, the theoretical person checks in the affirmative that he has mechanical ingenuity. In the division of peculiarities of people, he dislikes "bolshevists."

A few of the items which were more heavily weighted to characterize the non-theoretical individual are that he would like to be a "sales manager" and would have liked to have been "John Wanamaker, merchant." He dislikes the subject of "chemistry."

The economic man as described by Allport (6) is "characteristically interested in what is useful. Based originally upon the satisfaction of bodily needs (self-preservation), the interest in utilities develops to embrace the practical affairs of the business world. This type is thoroughly practical and conforms well to the prevailing stereotype of the average American business man."

On the "Economic" Scale of the *Strong Vocational Interest Blank*, some of the more heavily weighted items which appeared to differentiate the *economic* man from the *non economic* man indicate that he likes the occupation of "sales manager" and likes to "develop business systems." He would have liked being "Henry Ford, manufacturer", "J. P. Morgan, financier", or "John Wanamaker, merchant." This "typical business man" is indifferent toward the pastime of "reading a book" as compared with that of "going to the movies." He definitely dislikes the occupations of "artist" and "author of a novel", and has the same attitude toward the subject of "literature." He likewise records dislike for "absent-minded people", "socialists", and "bolshevists." He would

prefer "supervisory responsibility (head of a department of 300 people engaged in typical business operation)" to "technical responsibility."

In contrast, the non-economic man indicates that he likes such occupations as "clergyman", "college professor", "magazine writer", "poet", "school teacher", "sculptor", and "social worker." In the same vein, the non-economic person likes the subject "art" and likes "poetry." He admires "Luther Burbank, plant wizard", "Charles Dana Gibson, artist", and "Booth Tarkington, author." He likes amusements such as "observing birds", "visiting art galleries", and "museums", and listening to "symphony concerts."

In investigating the Bell *Adjustment Inventory*, the scales for "Home Adjustment", "Health Adjustment", and "Emotional Adjustment" on the Strong Blank showed relatively few weightings of two or more. This dearth of heavily weighted items and the large number of unit weight items in no apparent pattern, seems to indicate the inability of the Strong items to differentiate people scoring high from those scoring low on the Bell Home, Health, and Emotional Adjustment Scales. This fact is substantiated by the very low validity coefficient obtained.

Some of the more heavily weighted items which appear on the "social adjustment" scale that differentiate the *social* from the *non-social* type of person show that the former chose an occupation such as "consul." He liked subjects such as "dramatics", "literature", and "public speaking." The socially adjusted person indicated a liking for activities such as "interviewing clients", "opening a conversation with a stranger", "making a speech", "organizing a play", "meeting and directing people", "taking responsibility", "meeting new situations", and "entertaining others." He stated that he liked "quick tempered people" and that he "tells jokes well." He answered the following in the affirmative: "usually start activities of my group", "usually get other people to do what I want done", "usually liven up the group on a dull day", and "am quite sure of myself"

The non-social individual preferred to be a "member of a society" rather than an "officer in a society." He would rather "deal with things" than people. He would choose a "few intimate friends" rather than "many acquaintances." He would rather "listen to a story" than "tell a story." He stated that he "worries considerably about mistakes", his "feelings are easily hurt", and his "advice is practically never asked." The non-social person indicated a dislike for a "politician" and for "expressing judgments regardless of criticism."

An investigation of intelligence as measured by the American Council on Education *Psychological Examination* shows the following more heavily weighted items which appear to differentiate the person of *high intelligence* from the individual with a low intelligence rating. The former indicates a liking for the occupations of "author of a novel", "college professor", "editor", and "magazine writer." He dislikes the occupations of "life insurance salesman" and "office clerk." The person of high intelligence indicates a preference for "symphony concerts" and "poetry." He would rather read a book than go to a movie. He chooses the "Atlantic Monthly" for reading and prefers the school subjects of "algebra", "calculus", "geometry", "literature", and "philosophy." He enjoys "arguments", the "teaching of adults", and admires "independents in politics." He considers the most important factor affecting his work an "opportunity to make use of all knowledge and experience", and he says that he "can write a concise, well-organized report."

The occupation "floorwalker", the amusement "vaudeville", the subject "agriculture", and the man "John Wanamaker, merchant" are items which the individual of low intelligence rates as liked. He dislikes the occupations of "astronomer", "author of a technical book", "inventor", and the subject of "sociology." This dislike is also evidenced in the weighting of the items "Bolshevists", "writing personal letters", and "Booth Tarkington, author." The person with a low intelligence rating indicates that he would prefer to

"work for self in small business" than to "work in a large corporation with little chance of becoming president until age 55"; also, that he likes "many acquaintances" rather than a "few intimate friends."

Additional information was obtained on some of the new scales. A correlation was found between grade point averages for the first semester in school and the scores on the intelligence scale. The correlation was .42 ($N = 79$). The validity of this scale as a measure of scholastic aptitude is to be compared with a correlation of .39 between grade point averages and the scores on the American Council on Education *Psychological Examination* for the validating group ($N = 81$).

A correlation was found between the scores made with the Group II key (composed of mathematicians, engineers, chemists, and physicists) (17) of the *Strong Vocational Interest Blank* and the scores of the "theoretical" key. A correlation of .80 was obtained. This high correlation would seem to indicate that these two keys are to a large extent measuring approximately the same thing.

In order to determine whether the weightings on this new "economic" scale were similar to the Strong Key for Group VIII (accountant, office man, purchasing agent, banker) (17) a correlation was found for the scores of 89 subjects on the two keys. A correlation of .21 was obtained from these data, thus indicating that there is apparently little relationship between these keys. It would seem that the "economic" man is a different type of individual from the business man as represented by the accountant, office man, purchasing agent, banker group. The recent factor analysis study made by Ferguson, Humphreys, and Strong (11) shows how eight of the Strong scales (teacher, life insurance salesman, certified public accountant, office worker, physician, lawyer, Y. M. C. A. secretary, and chemist) are related to the six Allport-Vernon scales. The results indicate that the Strong scales used contain small factor loadings in the fifth factor. It is quite likely that the Strong "economic" scale of this study may produce heavier loadings and be an additional help for counseling in this area.

MEASURING PERSONALITY TRAITS

It seemed possible that a better estimation of the value of the "social adjustment" scale could be made by comparing resulting scores with social activity records. The 10 students of the validating group scoring highest on the new Strong "social adjustment" scale were compared with the ten students scoring lowest. A count was made of the number of activities in which these students participated during the semester the tests were taken. The students composing the "social" group belonged to a total of seven organizations, while those making up the "non-social" group belonged to only two.

Summary and Conclusion

The purpose of this study was to investigate the possibility of using the *Strong Vocational Interest Blank* to measure certain personality traits previously measurable only by the use of several other tests.

An analysis of the results showed that all of the scores based on the new keys were positively correlated with the traits measured. Validation coefficients based on a new group of subjects were: Bernreuter F1-C, .48; Bernreuter F2-S, .52; Allport-Vernon Theoretical, .48; Allport-Vernon Economic, .56; Bell Home Adjustment, .21; Bell Health Adjustment, .34; Bell Social Adjustment, .50; Bell Emotional Adjustment, .07; and the American Council on Education, Intelligence, .45. The reliabilities of these keys ranged from .70 for the Health Adjustment key to .87 for the Social Adjustment key.

In presenting the findings of this investigation, it should be remembered that the items of the *Strong Vocational Interest Blank* were not designed to be used as elements of a personality test. However, the weighted items of the Strong Blank give a fairly good picture of some of the factors making up a particular trait, e.g., Spranger's "economic" man (as measured by the Allport-Vernon test) and the weighted items on the new Strong "economic" key.

From the data gathered, the reliabilities indicate that the new Strong keys are fairly consistent in the material they are

measuring. It also appears that some traits can be measured with more accuracy by the Strong Blank than other traits. It does not seem advisable to continue work in the fields of "home adjustment", "health adjustment", or "emotional adjustment" with the *Strong Vocational Interest Blank* because of low validities obtained in these fields.

From this study, however, it does appear that a prediction of self-confidence and of sociability can be made with a fair amount of accuracy with the Strong Blank, and that types of individuals such as "theoretical" and "economic" (business man) can be determined fairly well by using the Strong keys for measurement of these traits. "Intelligence" scores based on responses to the Strong Blank show as high a correlation with college success as the American Council on Education *Psychological Examination* scores. This would indicate that in spite of relatively low validity coefficients when correlated with scores on the original tests, others of the new scales may be found to possess considerable practical validity in the evaluation of socially significant behavior.

REFERENCES

1. Allport, F. H. *Social Psychology*. New York: Houghton Mifflin Company, 1924. Chap. V, VI, XIV.
2. Allport, G. W. "A Test for Ascendancy-Submission", *Journal of Abnormal and Social Psychology*, XXIII, (1928), 118-136.
3. Allport, G. W. "Concepts of Trait and Personality", *Psychological Bulletin*, XXIV, (1927), 284-293.
4. Allport, G. W. *Personality*. New York: Henry Holt and Company, 1937. pp. 236-237, 303-304.
5. Allport, G. W. and Vernon, P. E. *Manual of Directions for a Study of Values*. Chicago: Houghton Mifflin Company, (1931).
6. *Manual for American Council on Education Psychological Examination*. Washington. American Council on Education, (1940).

MEASURING PERSONALITY TRAITS

7. Bell, Hugh M. *Manual for The Adjustment Inventory*, Stanford University Press, California, (1934).
8. Bernreuter, R. G. *Manual for The Personality Inventory*, Stanford University Press, California, (1938).
9. Chambers, O. R. "Character Trait Tests and Prognosis of College Achievement", *Journal of Abnormal and Social Psychology*, XX, (1925), 303-311.
10. Conklin, E. S. "Three Diagnostic Scorings for the Thurstone Personality Schedule", Indiana University Publications, Science Series, No. 6, (1937).
11. Ferguson, L. W., Humphreys, L. G., and Strong, F. W. "A Factorial Analysis of Interests and Values", *Journal of Educational Psychology*, XXXII (1941), 197-204.
12. Kelley, T. L. "The Scoring of Alternative Responses with Reference to Some Criterion", *Journal of Educational Psychology*, XXV, (1934), 504-510.
13. Kelly, E. L., Terman, L. M., and Miles, C. C. "Ability to Influence One's Score on a Typical Paper and Pencil Test of Personality", *Character and Personality*, IV, (1936), 206-215.
14. Kelly, E. L. "A Preliminary Analysis of Psychological Factors in Assortative Mating", *Psychological Bulletin*, XXXIV, (1937), 749.
15. Otis, A. S. *Manual for Self-Administering Tests on Mental Ability*, Yonkers-on-Hudson, New York: World Book Company, (1922).
16. Steinmetz, H. C. "Measuring Ability to Fake Occupational Interest", *Journal of Applied Psychology*, XVI, (1932), 123-130.
17. Strong, E. K., Jr. *Manual for Vocational Interest Blank for Men*, Stanford University Press, California, (1940).
18. Strong, E. K., Jr. "Procedure for Scoring an Interest Test", *Psychological Clinic*, XIX, (1930), 63-72.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

19. Strong, E. K., Jr. "Interest Maturity", *Personnel Journal*, XII, (1933), 77-90.
20. Strong, E. K., Jr. "Interests of Men and Women", *Journal of Social Psychology*, VII, (1936), 49-67.
21. Young, C. W. and Estabrooks, G. H. "Reports on the Young-Estabrooks 'Studiosness Scale' for Use with Strong Vocational Interest Blank for Men", *Journal of Educational Psychology*, XXVIII, (1937), 176-187.

A STUDY OF THE GENTRY VOCATIONAL INVENTORY

CLIFFORD FROELICH

State of North Dakota Occupational Information and Guidance Service

THE *Vocational Inventory* developed by Curtis G. Gentry, Director of Guidance and Secondary Education, Public Schools, Knoxville, Tennessee, contains 434 questions¹ Three hundred and eighty-four are in the *Inventory* proper, and the remainder constitute a personality inventory. The *Inventory* proper, according to Gentry's statement, "classifies the applicant's strengths and weaknesses with reference to these eight major groups"²: (1) social service, (2) literary work, (3) business, (4) law and government, (5) art, (6) mechanical designing, (7) mechanical construction, and (8) science. Gentry states in the *Manual of Directions* that the *Inventory* was constructed after "a fruitless search for an instrument which would yield a general vocational over-view of pupils and young adults."² The *Inventory* began to take shape in 1921 and has undergone many revisions since that time. It was copyrighted and published in its present form in 1940.

This study of the *Vocational Inventory* is based upon results obtained by administering it to 815 seniors in the high schools of Cass County, North Dakota. A majority (72 per cent) of the students were from a single school located in the metropolitan area of the county. In all cases the test was

¹C. G. Gentry. *Vocational Inventory*, (Nashville: Educational Test Bureau, 1940).

²C. G. Gentry. *Manual of Directions, Vocational Inventory*, (Nashville: Educational Test Bureau, 1941)

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

administered by persons specially trained in tests and measurements and was given under favorable testing conditions.

A statistical summary of scores of 1,000 Knoxville High School seniors is reported in the manual. Table 1 shows the total number of these students who received the highest scores in each of the groups as well as the percentages as reported by Gentry. The possible range of scores is from zero to 180. The highest score made by each student indicates the field for which greatest interest and ability are possessed.

TABLE 1
NUMBER OF HIGH SCORES BY GROUPS AS REPORTED BY GENTRY³

GROUP	I	II	III	IV	V	VI	VII	VIII
Total	168	135	143	184	99	23	172	76
Percent	16.8	13.5	14.3	18.4	9.9	2.3	17.2	7.6

A similar tabulation was made of the 815 students that were tested in Cass County. Table 2 gives the results of this tabulation. Separate tabulations are given by sexes.

TABLE 2
NUMBER OF HIGH SCORES BY GROUPS, 815 CASS COUNTY SENIORS

GROUP	I	II	III	IV	V	VI	VII	VIII
Girls Total	60	33	41	123	60	3	0	85
Percent	14.8	8.2	10.	30.4	14.8	0.7	0	21.
Boys Total	5	5	42	42	13	53	179	71
Percent	1.2	1.2	10.3	10.3	3.2	12.7	43.7	17.2
Total	65	38	83	165	73	56	179	156
Percent	7.9	4.6	10.2	20.2	8.9	6.9	21.9	19.4

It is evident from a comparison of girls and boys in Table 2 that there is a significant sex difference which must be taken into account in interpreting the results of this test although Gentry makes no statement in the *Manual* regarding sex differences. It is also interesting to note that when the scores of boys and girls are combined, there is still a wide discrepancy between the percentages as reported by Gentry and the percentages as found in the present study.

The question may be raised as to whether separate norms should be established for each sex in instances where sex differences are as pronounced as this study reveals. There is a

³Ibid, p 8

considerable difference of opinion on this point. Those who maintain that combined norms are justified base their opinion on the fact that the type and number of jobs open to women differs materially from those open to men. To them, establishing separate norms presumes that as many women as men would be advised to go into the field of mechanical construction, for example, when it is evident that many more men than women should be so advised. However, this position is not tenable for the writer. The establishment of separate norms would not necessarily mean that as many women as men would be guided into the mechanical construction field, but it would emphasize to the users of the inventory that sex differences do exist and that they must be taken into account. No instrument of this type can in any way supplant the necessity for every counselor's having a knowledge of the job demands and opportunities for both men and women in all occupational categories.

Question 383 of the *Inventory* asks the student to "name three vocations which appeal to you at the present time. Name your first choice first, if you have a first choice."¹ The writer noted that while using this instrument in a clinical situation there seemed to be a large proportion of the students whose claimed first choice of an occupation was in the same group as their highest score on the test. In order to check this relationship statistically, the student's personal choice was classified into one of the eight groups according to the "Psychological Classification of Occupations" as given by Gentry. The contingency coefficient for student's first choice and the highest score on the *Inventory* for 338 female high school seniors is .718. The *C* on a similar comparison for 300 male high school seniors is .733. When these two groups are combined, *C* equals .751. This correlation raises the question of whether the inventory, in a large measure, may not simply be a re-statement of the student's personal choice. An examination of the scattergrams for the correlations given above

¹C. G. Gentry *Op. cit.* p. 27.

indicates that 51 per cent of the cases fall on the axis, indicating that 51 per cent of the high school seniors have a personal vocational choice that is in the field of their highest score. Darley⁵ reports a maximum contingency coefficient of .5 between the claimed occupational interest types and measured occupational interest patterns for approximately 1,000 cases, with the Strong test, compared to .751 revealed by this study. Darley felt that he obtained a contingency coefficient as high as .5 only because he forced both the claimed and measured interest types into broad categories.

This close agreement between the scores on the *Inventory* and the student's stated choice may merely mean that he has, by a long and thorough process, come to such complete knowledge of occupational requirements that he reached the same conclusion that the inventory revealed in a comparatively short time. However, the individuals upon whom this study is based were members of school populations where no organized effort had been made to provide occupational information, and only a few of the students had related work experience. Furthermore, the counselors who interviewed the students following the administration of the *Inventory* found that in the majority of the cases the students had a very meager knowledge of the requirements of their stated vocational choices.

In the *Manual*, Gentry states, "An analysis of the *Inventory* returns on this sampling indicates that the second highest score is usually earned in an occupational group related to the first."⁶ Tables 3, 4, and 5 show the relationship of the highest score to the second highest score.

In some cases, Gentry's statement seems to be correct. However, in many cases it is difficult to find justification for his statement. For example, Group IV (Business), as shown in Table 5, indicates that 50 people earned their second score in Group III (Law and Government); 22 in Group II (Lit-

⁵J. G. Darley. *Clinical Aspects and Interpretation of the Strong Vocational Interest Blank* (New York: The Psychological Corporation, 1941).

⁶G. Gentry. *Op. cit.* p. 4

A STUDY OF THE GENTRY VOCATIONAL INVENTORY

TABLE 3
RELATIONSHIP OF HIGHEST TO SECOND HIGHEST SCORE
328 High School Females

Group in which student attained highest score										Total
Group		I	II	III	IV	V	VI	VII	VIII	
Group	I	0	8	5	17	7	0	0	25	62
in which	II	9	0	5	20	9	0	0	2	45
student	III	6	14	0	36	10	0	0	10	76
attained	IV	12	4	12	0	15	0	0	21	64
second	V	4	3	2	13	0	1	0	5	28
highest	VI	0	0	0	1	7	0	0	1	9
score	VII	0	0	0	0	0	0	0	2	2
	VIII	9	3	14	16	0	0	0	0	42
TOTAL		40	32	38	103	48	1	0	66	

TABLE 4
RELATIONSHIP OF HIGHEST TO SECOND HIGHEST SCORE
289 High School Males

Group in which student attained highest score										Total
Group		I	II	III	IV	V	VI	VII	VIII	
Group	I	0	0	1	0	0	0	0	1	2
in which	II	0	0	2	2	0	0	0	1	5
student	III	0	2	0	14	3	2	4	7	32
attained	IV	0	0	10	0	1	2	13	4	30
second	V	0	2	0	0	0	1	2	2	7
highest	VI	1	0	0	2	3	0	59	7	72
score	VII	1	0	10	12	2	36	0	28	89
	VIII	1	0	3	6	2	6	34	0	52
TOTAL		3	4	26	36	11	47	112	50	

TABLE 5
RELATIONSHIP OF HIGHEST TO SECOND HIGHEST SCORE
617 Males and Females Combined

Group in which student attained highest score										Total
Group		I	II	III	IV	V	VI	VII	VIII	
Group	I	0	8	6	17	7	0	0	26	64
in which	II	9	0	7	22	9	0	0	3	50
student	III	6	16	0	50	13	2	4	17	108
attained	IV	12	4	22	0	16	2	13	25	94
second	V	4	5	2	13	0	2	2	7	35
highest	VI	1	0	0	3	10	0	59	8	81
score	VII	1	0	10	12	2	36	0	30	91
	VIII	10	3	17	22	2	6	34	0	94
TOTAL		43	36	64	139	59	48	112	116	

erary); 22 in Group VIII (Science); 17 in Group I (Social Service); 13 in Group V (Art); and 12 in Group VII (Mechanical Construction). It appears that in 64 per cent of the cases the second highest scores will fall in some other group besides Group III. This is certainly a wider distribution than is indicated by the statement "usually earned in an occupational group related to the first."

The inference of Gentry's statement, it would seem, is that there are fairly high intercorrelations among certain scores if the second score is related to the first. Such overlapping, of course, is undesirable from a standpoint of measurement efficiency. The finding that such relationship is low therefore is a point in favor of the blank, even though it discredits Gentry's statement of close relationship.

In the *Manual of Directions*, Gentry assumes that interest will drive the person to acquire certain objective information or proficiency in the field of interest. According to Fryer's⁷ summary of the literature in the field, this so-called objective approach to the measurement of interest by the way of achievement or proficiency has never been unusually successful.

According to the statement of Gentry, the test is designed to measure the student's "strengths and weaknesses." Assuming that ability to deal with linguistic concepts and achievement in English are necessary requisites for success in the literary field, the following study was made. The Literary Group scores of 176 high school senior girls, selected at random, were compared with the scores made on the *Cooperative English Test* Form O.M., and the L-score of the American Council on Education *Psychological Examination*, 1938 Edition. The correlation between the English achievement score and the L-score was .73, with a P.E. of .022. A correlation of .74 with a P.E. of .018 was obtained by Beyers⁸ for these same measures on 500 college freshmen. The correlation between English achievement and the Literary Key on the Gentry *Inventory* was .589 with a P.E. of .033. The correlation was .605, with a P.E. of .055 between the L-score and the Literary Key. A partial correlation with the Gentry score held constant was .58. With the L-score held constant the correlation between the Literary Key and English achievement

⁷Douglas Fryer, *The Measurement of Interest* (New York: Henry Holt and Company, 1931).

⁸Otto Beyers, "Report of the Freshmen Testing Program, N. D. A. C., Fargo, North Dakota", 1940.

A STUDY OF THE GENTRY VOCATIONAL INVENTORY

was .095; with the English score held constant the correlation between the Literary Key and the L-score was .297.

These correlations seem to indicate that, to some extent, the Gentry test is measuring some factor which is not measured by the English and intelligence examinations. There is a question as to just what is being measured; possibly the high relationship between the personal choice and measured choice may offer one explanation. A common factor of experience in an academic high school may be another factor.

When intelligence is held constant, there seems to be little relationship between the English and the Gentry scores for the Literary Key. It is evident that the Literary Key gets at very little related to English achievement which is not already measured by an intelligence test. Yet in *The Student's Manual* that accompanies the test, the author states concerning this group that "one entering an occupation in the above group should like literature and English. He (or she) should be very much interested in composition, in writing articles, poems, themes, and reports."⁹

In making the above analysis to test the relationship of a single key to known objective measures, the writer does not contend that an interest inventory should show a marked positive relationship with such measures. On the contrary, an interest inventory can hardly be justified if it adds nothing to the scores of other measures. Though the findings here do not entirely discredit the usefulness of the *Vocational Inventory* as a measure of interest in the literary area, it does tend to disprove Gentry's claim that he is measuring "strengths and weaknesses."

While this study does not present conclusive evidence, the results at least indicate that this test should be more carefully standardized and evaluated before it is used in a counseling situation. The writer feels that the scores it now yields are of questionable value to the average guidance worker. Refine-

⁹C. G. Gentry. *Individual Analysis Report*. (Nashville: Educational Test Bureau, p. 6, 1940).

ment of procedures and further occupational standardization and follow-up data are necessary before the counselor can regard this inventory as a valid diagnostic tool.

THE RELATIONSHIP OF THE AFFECTIVE TOLERANCE INVENTORY TO OTHER PERSONALITY INVENTORIES

ROBERT I. WATSON
College of the City of New York

IN PRESENTING any new personality inventory it is important to establish the nature and extent of relationship that it bears to other scales designed to measure similar aspects of personality. The following pages present data on the relationship of the Watson-Fisher *Inventory of Affective Tolerance* to other standardized measures.¹

In constructing most inventories of emotional stability, items successfully used by others were usually selected from the clinical literature and combined with additional items for preliminary tryout with a more or less pragmatic outlook. If they "hung together" on internal validation and differentiated between extreme groups, a new inventory was constructed.

The *Inventory of Affective Tolerance* originated from a somewhat more theoretical framework. Items were collected with certain theoretical predilections as guiding principles. Statements of "symptom" were included in the preliminary tryout only if they seemed to be appropriate to this trait of affective tolerance. As a result, many so-called conventional items were not included. A brief description of this point of view follows.

An individual's affective tolerance is judged to be his ca-

¹This present report is considered as preliminary to a study by means of some factor technique.

capacity to handle his affective tensions; his capacity to adjust to affective disturbances. Among the chief aspects of affective tolerance are the capacities to withstand or endure emotional tension, to vent or discharge emotional tension, and to govern or direct emotional tension. It is the aggregate of these that the *Inventory of Affective Tolerance* purports to measure. A more complete discussion, including a description of the validation of the inventory, is given in an article by Fisher and Watson (3).

The Relationship to the Watson-Fisher Inventory of Affective Potency, to the Willoughby Personality Schedule, and to the Bernreuter Personal Inventory

The subjects employed in this section of the inquiry consisted of 55 boys and 97 girls, students at the University of Idaho, Southern Branch, who were described in connection with previous papers (3) (10). Besides the *Inventory of Affective Tolerance*, certain other measures including measures of general aptitude,² the Watson-Fisher *Inventory of Affective Potency*, the Willoughby *Personality Schedule*, and the Bernreuter *Personality Inventory* were administered to these students during the same semester.

The *Inventory of Affective Potency* purports to measure the trait of affective arousability, the strength and duration of our everyday affective responses (10). The Willoughby *Personality Schedule* is designed to measure "neurotic tendencies" (11). The Bernreuter *Personality Inventory* is designed to measure several aspects of personality. In view of the analyses of Flanagan (4) and Lorge (8), attention will be given to but two of these measures, designated by the symbols F1-C and F2-S in the inventory manual (2). The first is a measure of confidence in oneself. Persons scoring

²The correlations of the *Inventory of Affective Tolerance* with general aptitude were negligible. With the Otis *Self-Administering Test of Mental Ability, Higher Examinations*, the correlation coefficients were respectively, $-.04$ and $-.02$ for 50 boys and 90 girls. The correlations with the American Council on Education *Psychological Examination*, 1939 edition, were respectively, $.04$, $.03$, and $.04$ for L, Q, and Total scores in the case of the boys, and $.01$, $.10$, and $.06$ for the girls.

RELATIONSHIP OF THE AFFECTIVE TOLERANCE INVENTORY

low tend to be self-confident and well adjusted, and individuals scoring high tend to have feelings of inferiority. The second of these measures is one of sociability. At one extreme individuals tend to be non-social, at the other, gregarious.

TABLE 1

THE CORRELATION BETWEEN THE INVENTORY OF AFFECTIVE TOLERANCE AND THE WATSON-FISHER INVENTORY OF AFFECTIVE POTENCY, THE WILLOUGHBY PERSONALITY SCHEDULE, AND THE BERNREUTER PERSONALITY INVENTORY

	N	Tolerance					
		Boys			Girls		
		i	k	N	i	k	
Watson-Fisher	55	—14	.99	97	—21	.98	
Willoughby	47	—70	.71	94	—69	.73	
Bernreuter F1-C	46	—66	.75	87	—56	.83	
Bernreuter F2-S	46	—13	.99	87	—11	.99	

Table 1 gives the correlation between the scores on these inventories and scores on the tolerance inventory together with the corresponding coefficients of alienation. The correlations with the Willoughby and the F1-C or confidence factor of the Bernreuter test are substantial, ranging between —.56 and —.70. The correlations considerably exceed the minimum demanded at the one per cent level of significance (7). The correlations with the *Inventory of Affective Potency* and the sociability factor, F2-S, of the Bernreuter test are not significant at the one per cent level.

It would appear, then, that affective tolerance, as herein described and measured, bears considerable relation to self-confidence and lack of neurotic tendency, but little relation to sociability or affective potency.

This substantial relationship, however, should not be interpreted to mean that the tolerance inventory is so closely related to these other scales as to be superfluous for the measurement of individual differences. Consideration of the coefficients of alienation is pertinent in this connection. Reduction of the standard error of estimate by one-half, as expressed by a k of .50, requires that r be .866. The smallest coeffi-

cient of alienation found here is .71. Garrett (5) states that "For r 's of .80 or less the coefficients of alienation are clearly so large that predictions of individual scores based upon the regression equation are little better than a 'guess'." Although a substantial relation does exist between scores on the affective tolerance inventory and scores on these other inventories, it is evident that the *Inventory of Affective Tolerance* measures something other than whatever is measured by these two inventories.

The Relationship to the Colgate Personal Inventory C 2 and the Bell Adjustment Inventory

Fifty-nine white female student nurses, who were tested while receiving three months of their training at the Idaho State Mental Hospital (South) at Blackfoot, Idaho, took the *Inventory of Affective Tolerance*, the *Colgate Personal Inventory C2*, and the *Bell Adjustment Inventory*.³ The background of these subjects is summarized in Table 2.

TABLE 2
DESCRIPTION OF THE BACKGROUND OF FIFTY-NINE STUDENT NURSES

	Mean	σ	Range
Age	23.25	2.56	19-33
School Year Completed.	12.61	1.09	12-16
Months of Training Completed.	28.87	4.50	12-33
Otis <i>Self-Administering Test</i> Scores	45.90	9.18	25-65
Moss <i>Nursing Aptitude Test</i> Scores	139.90	30.02	41-186

The *Colgate Personal Inventory C 2* is designed to measure traits of introversion-extroversion. (6) The *Bell Adjustment Inventory* measures home, health, social, emotional, and occupational adjustment.¹ The higher the score in this inventory, the more unsatisfactory the adjustment.

The product-moment correlations between the *Inventory of Affective Tolerance* and the other inventories are presented in Table 3. All correlations are negative and all but two are significant at the one per cent level. (7) The correlation with the *Colgate Personal Inventory C 2* is not quite

³I wish to acknowledge the courtesy of Mr. Barney Bybee, then psychometrician at the Idaho State Mental Hospital, South, in supplying these scores.

RELATIONSHIP OF THE AFFECTIVE TOLERANCE INVENTORY

TABLE 3

THE CORRELATIONS BETWEEN THE INVENTORY OF AFFECTIVE TOLERANCE
AND THE COLGATE PERSONAL INVENTORY C 2 AND
THE BELL ADJUSTMENT INVENTORY

Measure	N	Tolerance	
		<i>r</i>	<i>k</i>
Colgate C-2	59	— .30	.95
Bell Home	59	— .40	.92
Bell Health	59	— .11	.99
Bell Social	59	— .60	.80
Bell Emotional	59	— .54	.84
Bell Vocational	45	— .50	.87
Bell Total ¹	59	— .56	.83

¹Vocational items are omitted since not all subjects took the form of the Bell Inventory that includes this section.

significant, and the correlation with the Bell Health Score is entirely negligible. Health adjustment and introversion-extroversion are apparently not significantly correlated with affective tolerance. The remaining correlations with Bell scores range from —.40 to —.60. Affective tolerance bears some relation to home, social, emotional, vocational, and total adjustment as measured by the Bell Inventory.

Although there is evidence that some degree of relationship exists, nevertheless the coefficients of alienation, which are also reported in Table 3, do not encourage the view that the two inventories can be used interchangeably. The smallest coefficient of alienation is .80, which implies 20 per cent efficiency in prediction.

Commonality of Items

An attempt was made to find out what items the Watson-Fisher inventory and the other personality measures had in common. Such an inquiry is pertinent in a field of investigation marked by repetition of items from inventory to inventory. Does the approach from the theoretical point of view earlier expressed result in the selection of the same items as the more pragmatic approach?

Two degrees of similarity of content can be distinguished. The first category adopted included items in which the con-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

tent apparently did not differ in scope, e. g., item 4 of the tolerance blank, "*I keep in the background at social gatherings*", as compared to item 118 of the Bernreuter, "*Do you keep in the background at social functions?*" In the second category were items in which the similarity was one of a whole to a part or a part to a whole, e. g., item 60 of the tolerance inventory, "*I control my feelings of grief or sorrow*", as compared to item 35 of the Bell, "*Are you easily moved to tears?*" The two categories will be referred to as "similar" and "partially similar," respectively.

The items of the tolerance inventory were checked against the four other personality measures used in one or another of the two samples studied for similar and partially similar items. The results of this subjective, somewhat crude analysis are presented in Table 4.

TABLE 4

THE SIMILARITY OF CONTENT OF ITEMS OF CERTAIN PERSONALITY MEASURES TO THE 61 ITEMS CONTAINED IN THE INVENTORY OF AFFECTIVE TOLERANCE.

Measure	Similar		Partially Similar		Similar or Partially Similar	
	N	%	N	%	N	%
Bernreuter	16	26	2	3	18	30
Willoughby	7	11	5	8	12	20
Laird C-2	6	10	4	7	10	16
Bell	14	23	4	7	18	30

It is apparent that similarity of content expressed in this fashion does occur. However, not more than 30 per cent of the items of the *Inventory of Affective Tolerance* appear in any one of these measures.

The commonality of items can be expressed in another fashion, namely, the total number of the 61 tolerance items appearing in one or more of the other measures as either a "similar" or a "partially similar" item. There were 27 such items. It is evident, then, that many of the items reported as similar in the data of Table 4 were found in two or more

of the other blanks. In all, 34 items or 56 per cent of the items in the tolerance inventory do not appear in the other four blanks. Apparently, then, there is no great similarity of items in the tolerance inventory and those contained in the other inventories studied.

Conclusions

1. Affective tolerance bears substantial relationship to confidence in oneself, lack of neurotic tendency, and social and emotional stability as measured by the scales used

2 Little or no relation is found between affective tolerance and affective potency, sociability, or health adjustment as measured by the scales used.

3. No relations are found to be of such a magnitude as to make the *Inventory of Affective Tolerance* superfluous, since the smallest coefficient of alienation is .71.

4. More than half the items contained in the *Inventory of Affective Tolerance* do not appear in the other personality measures employed.

REFERENCES

- 1 Bell, H. M. *Manual for the Adjustment Inventory: Adult Form*. Stanford University: Stanford University Press, 4 pp.
2. Bernreuter, R. G. *Manual for the Personality Inventory*. Stanford University: Stanford University Press, 6 pp.
3. Fisher, V. E. and Watson, R. I. "An Inventory of Affective Tolerance", *Journal of Psychology*, XII (1941), 149-157.
4. Flanagan, J. C. *Factor Analysis in the Study of Personality*. Stanford University: Stanford University Press, 1935 103 pp.
5. Garrett, H. E. *Statistics in Psychology and Education*. New York: Longmans, 1937. 493 pp.
6. Land, D. A. *General Information and Directions for Using the Colgate Tests of Emotional Outlets*. Hamilton: Hamilton Republican, 4 pp.
7. Lindquist, E. F. *Statistical Analysis in Educational Research*. Boston: Houghton-Mifflin, 1940. 266 pp.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

8. Lorge, I. "Personality Tests by Fiat. I The Analysis of the Total Trait Scores and Keys of the Bernreuter Personality Inventory", *Journal of Educational Psychology*, XXVI (1935), 273-278.
9. Otis, A. S. *Manual of Directions and Key: Otis Self-Administering Tests of Mental Ability*. Yonkers-on-Hudson: World Book, 12 pp.
10. Watson, R. I. and Fisher, V. E. "An Inventory of Affective Potency", *Journal of Psychology*, XII (1941), 139-148
11. Willoughby, R. R. *Directions: Willoughby (Clark-Thurstone) Personality Schedule* Providence: Author, 2 pp.

THE INFLUENCE OF TRAINING ON MECHANICAL APTITUDE TEST SCORES

RICHARD W. FAUBION and EARLE A. CLEVELAND

Air Corps Technical Training Command

and

THOMAS W. HARRELL

University of Illinois

THE present study is designed to investigate the influence of training on scores of the *Mechanical Movements* and *Surface Development Tests* (similar to many so-called aptitude tests), which look highly susceptible to training. In previous studies Harrell and Faubion found the *Mechanical Movements* and *Surface Development Tests* to be valid in predicting course grades of student airplane mechanics. The *Surface Development Test* correlated .55 with composite airplane mechanics grades for 84 men (1), and .47 with a slightly different composite for 105 other men (2). For the same two groups the correlations of the *Surface Development Test* with mechanical drafting and blueprint reading were .54 and .50, respectively. The *Mechanical Movements Test* correlated .39 and .26 with composite grades in the same two groups.

These tests developed by Thurstone are fairly familiar (3). The *Mechanical Movements Test* requires an individual to figure out how machine parts, particularly gears and pulleys, work. The *Surface Development Test* involves matching similar parts for drawings shown in two dimensions and in three-dimensional perspective.

The procedure was to study two groups which were matched for mental test scores but which differed in the

amount of mechanical training they had received. The two groups were each composed of 100 soldiers and were matched on the basis of scores on a mental test similar to the Henmon-Nelson. One group consisted of Air Corps recruits who had not as yet been entered in any of the technical courses offered by the Air Corps Technical Schools. The second group was composed of airplane mechanics students who had just finished a six-week basic training course in mechanical drafting and blueprint reading, elements of metalwork, elements of electricity, shop mathematics, and air corps fundamentals. The mean raw score for the untrained recruits on the mental test was 57.0, as compared with a mean raw score of 57.3 for the trained group. The difference between the two means is only one-fourth the standard error of the difference and consequently is quite insignificant. The variability of the two groups is also similar, a standard deviation of 8.96 being found for the recruits as against a standard deviation of 8.97 for the students.

The two groups were not intentionally paired for previous mechanical experience or training, but were selected at random from groups of Air Corps recruits and Air Corps Technical School students, respectively.

Special attention will be given to the mechanical drafting and blueprint reading course, as well as the elements of metalwork course, because of their apparent similarity to the tests. Mechanical drafting and blueprint reading, a forty-hour course, had the following outline:

- a. Fundamental principles of mechanical drafting.
- b. Exercises in orthographic projection.
- c. Development of surfaces.
- d. Blueprint reading.
- e. Exercises in blueprint reading.

Elements of metalwork, a sixty-hour course, had the following outline:

- a. Properties and uses of the common metals.

MECHANICAL APTITUDE TEST SCORES

- b. The care and use of the common tools needed in the repair and manufacture of small parts.
- c. Metalwork—projects in drilling, filing, thread cutting, reaming, etc
- d. Soldering—soft and hard soldering.
- e. Brazing.

The results given in Table 1 show no significant differences. The trained men had a score of 18.4 on the *Surface Development Test* as compared with 17.9 for the recruits. This difference is less than half the standard error of the difference.

TABLE 1

COMPARISON BETWEEN 100 RECRUITS AND 100 SOLDIERS TRAINED IN BASIC AIRPLANE MECHANICS

	Mean for Recruits	Mean for Trained Men	Difference	Sigma of Difference
Mental Test	57.0	57.3	0.3	1.2
Surface Development . . .	17.9	18.4	0.5	0.9
Mechanical Movements (No. Right)	32.2	32.5	0.3	1.2
Mechanical Movements (Rt. —W)	18.2	17.6	0.6	1.7

The *Mechanical Movements Test* scores were treated in two ways: (1) Number right and (2) Rights minus wrongs. Since the number of choices varies from one question to another, the use of a simple correction formula would be questionable. The trained group had a mean score of .3 of a point higher for the number right. The recruits had a mean score .6 of a point higher when the correction formula was used. Neither of these differences is as large as half the standard error of the difference.

These results indicate that six weeks of intensive training in mechanical courses do not significantly increase mechanical aptitude test scores, even where the test is very similar to the activities carried out in the training. This is strikingly true of the *Surface Development Test*, in which the items resemble mechanical drafting and blueprint reading work. No conclusion can be drawn as to how far this result can be generalized; possibly longer training or earlier training would show a

significant increase. The authors wish to point out, however, that the present results are contrary to statements often made about a mechanical aptitude test.

REFERENCES

1. Harrell, T. W. and Faubion, R. W. "Selection Tests for Aviation Mechanics", *Journal of Consulting Psychology*, IV, (1940), 104-105.
2. ——— "Primary Mental Abilities and Aviation Maintenance Courses", *Educational and Psychological Measurement*, I, (1941), 59-66.
3. Thurstone, L. L. *Primary Mental Abilities*. Psychometric Monographs, No. 1. Chicago. University of Chicago Press, 1938. 128pp.

NEW TESTS[†]

American School Achievement Tests, by Robert V. Young, Willis E. Pratt, and Frank Gatto. 1941. Forms A and B. Primary Battery I, for grade 1. Time, 35 minutes. \$2.50 per 100; 3c each; specimen set 25c. Primary Battery II, for grades 2 and 3. Time, 85 minutes. \$4.00 per 100; 5c each; specimen set 30c. Published by the Public School Publishing Company, 509-513 North East Street, Bloomington, Illinois.

American School Reading Readiness Test, by Robert V. Young, Willis E. Pratt, and Carroll A. Whitmer. For kindergarten and grade 1. Time, about 30 minutes. Form A, \$4.00 per 100; 5c each; specimen set 25c. Published by the Public School Publishing Company, 509-513 North East Street, Bloomington, Illinois.

Arithmetical Reasoning Test, by Alfred J. Cardall. 1941. For academic and technical prediction. For 12th grade level and above. Time, 40 minutes. Forms A and B, 5c each; specimen set 20c. Published by Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois.

Chicago Tests for Primary Mental Abilities, by L. L. Thurstone and Thelma Gwinn Thurstone. 1942. For ages 11 to 17. Time, 40 minutes for each of six booklets. \$5.00 per 25 sets; \$9.00 per 50 sets; \$15.00 per 100 sets; \$70.00 per 500 sets; specimen set \$1.00; supplementary supplies additional. Published by the American Council on Education, 744 Jackson Place, Washington, D. C.

College English Test, National Achievement Tests, by A. C. Jordan. 1941. For high school seniors and college freshmen. Time, about 45 minutes. Forms A and B, \$2.50 per 25; 100 or more copies 7½c each. Published by the Acorn Publishing Company, Rockville Centre, Long Island, New York.

[†]Prepared by Jane Gilbert.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Detroit Alpha Intelligence Test, by Harry J. Baker. 1941. For grades 4 to 8. Time, 32 minutes. Forms S and T, \$3.50 per 100; 4c each; specimen set 15c. Published by the Public School Publishing Company, 509-513 North East Street, Bloomington, Illinois.

English Minimum Essentials Test, by J. C. Triessler. Revised 1941. For grades 8 to 12. Time, about 40 minutes. Forms A, B, and C, 75c per 25; 4c each; specimen set 10c. Published by the Public School Publishing Company, 509-513 North East Street, Bloomington, Illinois.

Every-Day Life, by Leland H. Stott. 1941. To measure three factors in self-reliance. For high school students. Time, about 30 minutes. Hand- or machine-scored. \$4.00 per 100; \$2.25 per 50; specimen set 15c; machine-scoring answer sheets, \$2.00 per 100 up to 500. Published by the Sheridan Supply Company, P. O. Box 837, Beverly Hills, California.

Furbay-Schrammel Social Comprehension Test, by John H. Furbay and H. E. Schrammel. 1941. For high school and college students, and adults. Time, 80 minutes. Form A, \$1.70 per 25; 7c each; specimen set 15c. Published by the Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas.

Interest Inventory for Elementary Grades, by Mitchell Dreese and Elizabeth Mooney. 1941. For grades 4, 5, and 6. Time, about 30 minutes. Published by Center for Psychological Service, George Washington University, Washington, D. C.

Inventory of Social Behavior, by Ellis Weitzman. 1941. For ages 16 to 25. Time, about 20 minutes. \$4.00 per 100; \$2.25 per 50; specimen set 15c. Published by the Sheridan Supply Company, P. O. Box 837, Beverly Hills, California.

Iowa Every-Pupil Tests of Basic Skills—Form M, by H. F. Spitzer, Ernest Horn, Maude McBroom, H. A. Greene, and E. F. Lindquist. 1941. Test A, Silent Reading Comprehension; Test B, Work-Study Skills; Test C, Basic

NEW TESTS

Language Skills; Test D, Basic Arithmetic Skills. Elementary Battery for grades 3 to 5. Time, about 85 minutes. \$1.15 per 25 for each test; \$3.75 per 25 for complete battery. Advanced Battery for grades 5 to 8. \$1.25 per 25 for each test; \$4.00 per 25 for complete battery. Published by Houghton Mifflin Company, 2 Park Street, Boston, Massachusetts.

Iowa Placement Examinations, English Training—Form M, constructed by M. F. Carpenter, G. D. Stoddard, and L. W. Miller; revised by M. F. Carpenter and D. B. Stuit. Revised 1941. For college students. Time, 45 minutes. Hand- and machine-scored. \$4.00 per 100; machine-scoring answer sheets 1½c each; specimen set 20c. Published by the Bureau of Educational Research and Service, State University of Iowa, Iowa City, Iowa.

Iowa Placement Examinations, Foreign Language Aptitude—Form M, constructed by G. D. Stoddard; revised by Grace Cochran, J. R. Nielson, and D. B. Stuit. Revised 1941. For college students. Time, 45 minutes. Hand- and machine-scored. \$4.00 per 100; machine-scoring answer sheets 2c; specimen set 20c. Published by the Bureau of Educational Research and Service, State University of Iowa, Iowa City, Iowa

Iowa Placement Examinations, Physics Aptitude—Form M, constructed by G. D. Stoddard and C. J. Lapp; revised by C. J. Lapp and D. B. Stuit. Revised 1941. For college students. Time, 50 minutes. Hand- and machine-scored. \$4.00 per 100; machine-scoring answer sheets 1½c each; specimen set 20c. Published by the Bureau of Educational Research and Service, State University of Iowa, Iowa City, Iowa.

Kansas Spelling Test, by H. E. Schrammel, O. M. Rasmussen, and Wayne Gordon. 1941. Test I for grades 1 to 3; Test II for grades 4 to 6; Test III for grades 7 to 9. Time, 15 minutes. Forms A and B, 50c per 25; specimen set 15c. Published by the Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Language Essentials Tests, by Vera Davis and H. E. Schrammel. 1941. For grades 4 to 8. Time, 30 minutes. Forms A and B, \$1.00 per 25; \$4.00 per 100; \$36.00 per 1000; specimen set 25c. Published by the Educational Test Bureau, Minneapolis, Minnesota.

McDougal General Science Test, by Clyde R. McDougal. 1941. For high school students. Time, 40 minutes each for Test I and Test II. 50c per 25; 2½c each; specimen set 15c. Published by the Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas.

Primary Business Interests, by Alfred J. Cardall. 1941. For high school, college and adult levels. Time, about 20 minutes. Hand- or machine-scored. 5c each; scoring keys 25c; specimen set 35c. Published by Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois.

Primary Reading Tests, by Albert G. Reilley. 1941. For grade 1. Form B, 85c per 25. Published by Houghton Mifflin Company, 2 Park Street, Boston, Massachusetts.

Recreation Inquiry, by Richard Wilkinson and Sidney L. Pressey. For high school and college students. Time, about 50 minutes. \$1.00 per 25; \$3.00 per 100; specimen set 15c. Published by the Psychological Corporation, 522 Fifth Avenue, New York City.

Wiksell-Filkin Library Instructional Tests, by Wesley Wiksell and Mary Filkin. 1941. For high school and college students. 25 separate tests each on a different subject. \$3.75 per 25 complete batteries. Published by the Acorn Publishing Company, Rockville Centre, Long Island, New York.

Wilson Scales of Stability and of Instability, by Matthew H. Wilson. 1941. For junior and senior high school and college students, and adults. Time, 20 to 30 minutes. \$1.15 per 25; 5c each; specimen set 15c. Published by the Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas.

MEASUREMENT ABSTRACTS*

Baxter, Brent. "An Experimental Analysis of the Contributions of Speed and Level in an Intelligence Test." *Journal of Educational Psychology*, XXXII (1941), 285-96

Three measurements were taken of performance on an intelligence test: speed, the time to complete the entire test; power, the number of items correct at the end of a given time; and level, the number of items correct with unlimited time. Speed and level are uncorrelated. "Speed and level contribute the entire variance of power." Level is more important than speed in determining college grades. When the students are tested individually, prediction (of Army Alpha scores, of college aptitude test scores, and of college grades) "through the combination of speed and level in multiple correlation is greater than that possible" with the more usual scores of power. When the students are tested in groups, this superiority vanishes. *H M. Wolfe.*

Carroll, John B. "A Factor Analysis of Verbal Abilities." *Psychometrika*, VI (1941), 279-308.

A multiple-factor analysis was made of a battery of 42 tests of verbal abilities administered to 119 college adults. Where necessary, the distributions of test scores were normalized before the inter-test correlations were computed. Thurstone's M (Memory or Rote Learning) factor has been confirmed, but his V (Verbal Relations) factor seems to have been split into two or possibly three factors, C, J, and G. I is W (Word Fluency) factor has been split into two factors, A and E. The C factor seems to represent the richness of the individual's stock of linguistic responses, and the J factor seems to involve the ability to handle semantic relationships. No satisfactory interpretation can as yet be made of the G factor. The A factor seems to correspond to the speed of association for common words where there is a high degree of restriction as to appropriate responses. The E factor is described as an associational facility with verbal material where

*Edited by Professor Forrest A. Kingsbury.

the only restriction is that the responses must be syntactically coherent. The new factors are: I', facility and fluency in oral speech; H, facility in attaching appropriate names or symbols to stimuli; and D, speed of articulatory movements. (Courtesy *Psychometrika*.)

Chapanis, Alphonse. "Notes on the Rapid Calculation of Item Validities." *Journal of Educational Psychology*, XXXII (1941), 297-304.

Several shortcuts in the estimation of item validities by means of the biserial correlation coefficient are suggested. When one is not interested in inter-test comparisons, the constant $\frac{1}{\sigma_T}$ may be eliminated. By employing the same class interval and assumed mean i may be omitted and the means in the formula replaced by deviations of guessed means. Moreover, if the range of abilities tested is homogeneous, z is unnecessary. Formulae facilitating transformation of reduced coefficients are presented in case inter-test comparisons are later desired. *V. Brown*.

Crissy, William J. E. "A Reply to an Examinee's Reactions to the National Teacher Examinations." *Journal of Higher Education*, XII (1941), 484-487.

This article takes each of the particular criticisms in turn and indicates how each problem of test construction was handled in compiling the National Teacher Examinations. The advantages of the particular test form used are listed and year-to-year comparability of test scores is claimed. The construction of test items and scoring key is briefly outlined. The purpose of the examinations and the use of the test results are considered and precautions taken in this area are pointed out. *D. A. Peterson*.

Ferguson, George A. "The Factorial Interpretation of Test Difficulty." *Psychometrika*, VI (1941), 323-30.

This paper discusses the influence of test difficulty on the correlation between test items and between tests. The greater the difference in difficulty between two test items or between two tests, the smaller the maximum correlation between them. In general, the greater the number of degrees of difficulty among the items in a test or among the tests in a battery, the

higher the rank of the matrix of intercorrelations; that is, differences in difficulty are represented in the factorial configuration as additional factors. The author suggests that if all tests included in a battery are roughly homogenous with respect to difficulty, existing hierarchies will be more clearly defined and psychological interpretation will be more meaningful. (Courtesy *Psychometrika*.)

Flanagan, John C. "A Preliminary Study of the Validity of the 1940 Edition of the National Teacher Examinations." *School and Society*, LIV (1941), 59-64.

Because conclusive validation is not yet possible, this report aims to do no more than to review evidence now available respecting the 1940 National Teacher Examinations. Two meanings of "validity" are distinguished: (1) Do the tests satisfactorily get at the content and mental processes indicated in the outline and specifications by which they were planned? (2) Do they aid in distinguishing between better and poorer teachers as measured by any of numerous criteria of teacher excellence? The former is partially answered by the agreement of the 10 or 12 cooperating experts who critically examined the tests and their specifications, and by intercorrelations between tests of the battery; the latter, by citation of several lines of evidence. One line is student ratings on 49 teachers in 22 systems (at least 2 in each system) who had taken these tests and were chosen so as to reveal considerable spread of scores on the "common examinations." Correlation between ratings and test-scores was .51. Supervisors' ratings on these teachers on a number of items are also cited. The five highest of these correlated around .50 with test-scores. In general, the study shows that the examinations have some predictive value as to the teacher's general effectiveness and desirability, and also points out other significant items. *F. A. Kingsbury*.

Froehlich, Gustav J. "A Simple Index of Test Reliability." *Journal of Educational Psychology*, XXXII (1941), 381-85

A simple adaptation of the Kuder-Richardson index of test reliability is described, namely:

$$r = \frac{\sigma^2 n - M (n - M)}{\sigma^2 (n - 1)}$$

Since this formula involves only the number of items in the test, the mean of the test scores, and their standard deviation, it is offered as an index of test reliability easily applied by teachers and others who are limited with respect to time and statistical background. An empirical check, using the Wisconsin Achievement Test on some 2000 individuals, shows that reliability coefficients on the total battery and five parts, as computed by this formula, run slightly lower (.017 to .058) than Spearman-Brown r 's, the two rank orders of the six r 's being identical. *F. A. Kingsbury.*

Gritten, Frances and Johnson, Donald M. "Individual Differences in Judging Multiple-Choice Questions." *Journal of Educational Psychology*, XXXII (1941), 423-430.

Form A of the *Nelson-Denny Vocabulary Test* was given with instructions not to guess; form B with instructions to answer all questions and to rate each judgment on a confidence scale. Four different achievement scores from Form A were correlated with the confidence and achievement scores from Form B. The results indicated that with instructions not to guess, the more confident subjects will attempt and correctly answer more items, and that the conventional formula, $R = \frac{W}{n-1}$, could properly be called a correction for individual differences in confidence. *V. Brown.*

Haggerty, Lida Harmer. "An Empirical Evaluation of the Accomplishment Quotient A Four Year Study at the Junior High School Level." *Journal of Experimental Education*, X (1941), 78-90.

"The AQ is a distinctly unreliable measure." This conclusion was reached after studying data for 163 subjects over a four-year period. Intelligence was measured eight times, using four different tests, and achievement four times, combining Forms V and W of the New Stanford Achievement Test. "There is a mean inter- r of only .35 for the eight AQ distributions," in spite of duplication among the measures. "Large numbers of pupils (seem to) achieve up to capacity or fall below capacity without the slightest change in their actual work by merely changing from one accepted test to another in measuring intelligence."

A composite of all achievement scores correlates .94 with a composite of all intelligence scores *H. M. Wolfe*.

Hartmann, George W. "A Critique of the Common Method of Estimating Vocabulary Size, Together with Some Data on the Absolute Word Knowledge of Educated Adults" *Journal of Educational Psychology*, XXXII (1941), 351-358.

Vocabulary estimates based upon samples of varying size drawn from the same dictionary were found to be fairly stable, although results indicated that less than fifty words do not yield an accurate measure. When samples chosen from dictionaries of varying size were compared, vocabulary estimates were discovered to be dependent upon the size of the dictionary. Commonly accepted estimates need upward revision, for the present study demonstrated that the recognition vocabulary of the average undergraduate is in excess of 200,000 words *V. Brown*.

Horst, Paul, and collaborators. *The Prediction of Personal Adjustment*. New York: Social Science Research Council, pp. xii+455. 1941.

This monograph is a study of the logic and methodology of the prediction of personal adjustment, prepared under the supervision of the Committee on Social Adjustment of the Social Science Research Council. It is oriented primarily with studies in prediction of adjustments in four fields, namely, school success, vocation, marriage, and crime, with a supplementary memorandum on problems of prediction in the national defense program. Five supplementary studies by collaborators are included dealing with case-study techniques, mathematical and tabulation techniques, reduction of number of variables (factor grouping), combining and weighting measures, and five mathematical problems. In the systematic section, detailed descriptions of tests and other instruments are omitted, and the major methodological aspects of the prediction problems are summarized and analyzed in order. A chapter is devoted to suggestions for research projects in the prediction of individual behavior. *F. A. Kingsbury*.

Jackson, R. W. B. "Some Difficulties in the Application of the Analysis of Covariance Method to Educational Problems." *Journal of Educational Psychology*, XXXII (1941), 414-422.

Since the analysis of covariance is a statistical method developed mainly for use in another field, it seems inadvisable to apply it without question to educational problems. Although the method has been found very useful, it may be necessary to modify it slightly. Examples are given to illustrate some of the difficulties likely to be encountered in the adoption of this method and to demonstrate their possible solutions. *V. Brown.*

Langsam, Rosalind Streep. "A Factorial Analysis of Reading Ability." *Journal of Experimental Education*, X (1941), 57-63.

The factors involved in reading ability were determined using Thurstone's centroid method with rotation of axes. Twenty different subtests from reading and intelligence tests and one from the *Primary Mental Abilities* battery were analyzed. Three of the four factors found in reading tests were identified as being similar in character to three of Thurstone's primary abilities: a verbal factor V, "an ability to deal with verbal material"; a perceptual factor P, which in this material shows up as speed in "perceiving and selecting the correct word from other words offered as possible answers"; and a word factor W, "a fluency in dealing with words." A more tentative factor was "that of seeing relationships." *H. M. Wolfe.*

Lennon, Roger T. "Note on Line of Relation Method of Establishing Age or Grade Norms." *Journal of Educational Psychology*, XXXII (1941), 389-90.

Two methods of establishing age or grade norms are: (1) to find mean scores for successive age or grade groups and pass a norm line through them; (2) to determine empirically the correspondence of scores on the new test with those on a test whose "line of relation" has already been established and interpolate norms on such a line of relation. The author points out the condition under which the two methods yield identical results; namely, when the correlation of scores on each of the

two tests with age is the same. The correspondence method is applicable only when this condition is known to be satisfied.
F. A. Kingsbury.

McCormick, Thomas Carson. *Elementary Social Statistics*. New York: McGraw-Hill Book Company pp. 353. 1941

This elementary statistics textbook is designed primarily for students and workers in sociology rather than in psychology and education. The first part of the book deals with the nature and control of statistical inquiry. The second part is devoted to common statistical procedures: tabulation of distributions, graphs, measures of deviation, correlation techniques, sampling and sampling errors, the significance of differences, and analysis of time series. Statistical tables are also included at the end of the book. *Jane Gilbert.*

McNamara, John Joseph. "A New Method for Testing Advertising Effectiveness Through Eye Movement Photography." *Psychological Record*, IV (1941), 399-460.

In order to test advertising effectiveness one group of readers was asked to leaf through a magazine containing advertising matter. Their eye-movements were photographed with the Purdue Eye-Camera and the mean time spent on each part of an advertisement was recorded. The magazine was advance copy so the readers had no opportunity to see the copy prior to the experiment. Another group was given advertisements which had been cut up into parts and pasted on cardboard in heterogenous order. The length of time the reader required to identify the parts with the whole advertisement was recorded. The probability that the reader would look at the advertisement long enough to identify the advertiser was also computed. The reliability of these techniques was high. The correlation between mean time and probability scores was .48 for combined groups. The effect of magazine position, position on the spread, and cartoons on advertising effectiveness was also determined. *Jane Gilbert.*

Peters, Charles C. and Van Voorhis, Walter R. *Statistical Procedures and Their Mathematical Bases*. New York: McGraw-Hill Book Company. pp. 516. 1941.

This textbook is designed to explain the mathematical origins of statistical formulae in terms that can be understood

by statistical workers having little mathematical background. Many of the Fisher techniques are discussed in relation to theoretical statistics, although the limitations of these techniques as applied in the psychological and social sciences are also indicated. The authors present a brief discussion of calculus and elementary statistical procedures, measures of central tendency, variability, reliability, probability, multiple-factor analysis, curve fitting, partial and multiple correlation, the nature of Chi squared, and the techniques used in controlled experimentation. *Jane Gilbert.*

Remmers, H. H. and House, J. Milton. "Reliability of Multiple-choice Measuring Instruments as a Function of the Spearman-Brown Prophecy Formula, IV" *Journal of Educational Psychology*, XXXII (1941), 372-76.

The hypothesis tested is that the relation between changes in reliability of multiple-choice test-items and changes in number of response alternatives per test-item is predictable by the Spearman-Brown formula. A 60-item, five-alternative, multiple-choice arithmetic test was given to 771 junior high school pupils. Three derivative forms were constructed from this, having respectively four, three, and two alternative answers per item, the eliminated answers having been selected by lot. Four groups, equated as to I.Q., took separate forms of the test. Reliability coefficients of half-test (odd-even) and whole-test both showed regular decrease for the four forms with decreasing number of alternatives, thus supporting the hypothesis within this range of two to five alternative responses. *F. A. Kingsbury.*

Remmers, H. H. and Sageser, H. W. "Reliability of Multiple-choice Measuring Instruments as a Function of the Spearman-Brown Formula, V." *Journal of Educational Psychology*, XXXII (1941), 445-51.

The hypothesis tested is the same as that described in the Remmers and House article, abstracted above. Two equivalent attitude-scales (Remmers & Bues) of 37 items were combined and used in testing attitudes of agreement or disagreement on each of two college practices. Four derivative sets were prepared, providing respectively two, three, five, and

seven degrees of choice. From 87 to 112 university students filled out each of these forms. Each of the two tests was scored twice, once with equal values for all statements, once with items weighted by scale-value. Dividing each of the four sets of scores into the two original 37-item scales, four correlations were found for each set of papers. Corrected for skewness by being transformed into "z" functions, the obtained reliabilities with unweighted scores did not support the hypothesis; but when weighted in terms of the experimentally determined scale values of the scale-items, the data supported the hypothesis. *F. A. Kingsbury*

Ruch, Floyd L. "The Problem of Measuring Morale." *Journal of Educational Sociology*, IV (1941), 221-228.

At the present time one of the more effective tools developed to give an orderly description of public response is the opinion poll. The two basic problems in public opinion polling are: to get a representative sample, and to get the desired information from every case in the sample. The problem of sampling is discussed and references are given for specific techniques in the field. In the second problem basic types of defects are listed which lower the dependability and accuracy of questions. In conclusion possible additions to the public opinion poll technique in the measurement of morale are discussed. *D. A. Peterson.*

Satterthwaite, Franklin E. "Synthesis of Variance." *Psychometrika*, VI (1941), 309-16

The distribution of a linear combination of two statistics distributed as is Chi-square is studied. The degree of approximation involved in assuming a Chi-square distribution is illustrated for several representative cases. It is concluded that the approximation is sufficiently accurate to use in many practical applications. Illustrations are given of its use in extending the Chi-square, the Student "t" and the Fisher "z" tests to a wider range of problems. (Courtesy *Psychometrika*.)

Spache, George. "Deriving Comprehension, Rate and Accuracy of Reading Norms for a Short form of the

Metropolitan Achievement Reading Test." *Journal of Educational Psychology*, XXXII (1941), 359-64.

The *Metropolitan Achievement Tests*, although widely used, require a long testing time (about 2 $\frac{3}{4}$ hours for Intermediate Partial). The uses to which the test is put indicate that abbreviation of certain tests is preferable to omission of sub-tests, if the time has to be shortened. The method used in abbreviating the Reading Test is described, and correlation coefficients between total grade-score and shortened test raw-scores are cited. Grade-score norms for the short form are given, derived from the regression equations, and also percentile norms for reading accuracy, the latter based on private schools and of uncertain validity for public schools. Validity coefficients for the shortened form are found to be almost as high as the reliability coefficients of the long form. *F. A. Kingsbury*.

Thornton, G. R. "The Use of Tests of Persistence in the Prediction of Scholastic Achievement." *Journal of Educational Psychology*, XXXII (1941), 266-274.

A factor analysis of persistence tests revealed two unrelated factors; one appeared in the shock and pressure tests, the other in the word building and perceptual ability tests. The hypothesis that personality tests have value for predicting scholastic success in proportion to the degree of similarity between tests and classroom situations is suggested to explain the lower correlation between achievement and persistence found in this investigation. A formula utilizing scholastic efficiency and aspiration displayed in a previous school is presented for prediction of grades in a new school. *V. Brown*.

Travers, L. B. "Improving Practical Tests." *Personnel Journal*, XX (1941), 129-33.

This article discusses the advantages of evaluating the personal characteristics of testees while they are undergoing a practical test, rather than while interviewing them. The ratings under actual working conditions are claimed to be more reliable. Sample charts are given for a carpenter's practical test and the rating sheet used with it. *H. M. Wolfe*.

Traxler, Arthur E. "A Study of the Junior Scholastic Aptitude Test." *Journal of Educational Research*, XXXV (1941), 16-27.

The Junior Scholastic Aptitude Test consists of three parts: verbal section, containing five subtests; numerical section, containing three subtests; and an experimental section, not included in pupil's score. A practice booklet is issued several days prior to the examination for the student to work through at his leisure. The administration of the test proper is on a secret basis. Results are reported to the school in terms of a derived score. The reliability, validity, and prognostic value of the test are discussed as indicated by correlations with other tests of academic aptitude, achievement tests, and school marks. "The data are not extensive enough to be conclusive, but it is hoped that they will be of some assistance in appraising and using this new test" *D. A. Peterson.*

Wherry, Robert J. "An Extension of the Doolittle Method to Simple Regression Problems." *Journal of Educational Psychology*, XXXII (1941), 459-464.

This article describes a new method for solving simple regression constants which the author has found very successful in teaching beginners. It is shorter not only because it involves fewer arithmetical operations, but also because it is more systematic. Other advantages claimed are that the checks are more certain and convincing, and once the beginner has mastered the technique involved in simple correlation, he is immediately able to solve multiple correlation constants in precisely the same manner with little further training. *V. Brown.*

Winetrout, Kenneth. "The National Teacher Examinations, 1941." *Journal of Higher Education*, XII (1941), 479-484.

The writer gives a brief, general description of the length of the examinations and the method of giving them. This is followed by criticisms both adverse and appreciative of the construction of the examinations and potential use of test results obtained. It is suggested that if the question form were

varied in the long testing periods (the examinations, total duration being twelve hours, are restricted to the use of the multiple-choice question form), a more adequate examination program might be obtained. The author believes that one section was concerned with measurement of attitudes rather than capacities and would substitute a three-point (conservative, liberal, radical) rating scale for the right-wrong classification used at present. Wording of questions, sectional influence, and factual emphasis are also discussed by the writer, who recently took the examinations. *D. A. Peterson.*

Young, Gale. "A Note on Multidimensional Psychophysical Analysis." *Psychometrika*, VI (1941), 331-33.

On viewing Thurstone's psychophysical scale from the point of view of the mathematical theory of one-parameter continuous groups, it appears that a variety of different psychological or statistical assumptions can all be made to lead to a scale possessing similar properties, though requiring different computational techniques for their determination. The natural extension to multi-dimensional scaling is indicated. (Courtesy *Psychometrika*)

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Volume II

APRIL, 1942

Number 2

A TEST FOR PRIMARY BUSINESS INTERESTS BASED ON A FUNCTIONAL OCCUPATIONAL CLASSIFICATION ..	113
--------------------------------------------------------------------------------------------	-----

Alfred J. Cardall

MEASUREMENT IN RURAL HOUSING

A PRELIMINARY REPORT ...	139
--------------------------	-----

Charles I. Mosier

PROCEDURE FOR HANDLING TESTS AND EXAMINATIONS.....	153
----------------------------------------------------	-----

John V. McQuitty

MACHINES IN CIVIL SERVICE TESTING	167
----------------------------------------	-----

Sidney W. Koran

PREDICTIVE VALUE OF CERTAIN "LAW APTITUDE" TESTS.....	201
-------------------------------------------------------	-----

E. L. Welker and T. W. Harrell

AN EXPLORATORY STUDY OF SOCIAL GUIDANCE AT THE COLLEGE LEVEL	209
--------------------------------------------------------------------	-----

Margaret Glockler Aldrich

NEW TESTS	217
-----------------	-----

MEASUREMENT ABSTRACTS	220
-----------------------------	-----

MEASUREMENT NEWS	227
------------------------	-----

Copyright, 1942, by
SCIENCE RESEARCH ASSOCIATES

PRINTED IN THE UNITED STATES OF AMERICA

A TEST FOR PRIMARY BUSINESS INTERESTS BASED ON A FUNCTIONAL OCCUPATIONAL CLASSIFICATION

ALFRED J. CARDALL
Boston University

AS VOCATIONAL GUIDANCE leaves the area of pleasant gestures and advances towards a realistic process, it becomes more and more dependent upon better methods of evaluating an individual's interests and potentialities. In spite of many imperfections, psychological tests still constitute our best means of diagnosis. It is questionable, however, if even the best analysis of an individual's aptitudes, special abilities, and personality traits helps materially in indicating an occupational area in which the individual will find stimulation, economic independence, and satisfaction in his work. The results of such tests in the hands of a careful counselor serve chiefly in the determination of the "vocational risk" involved in the pursuit of those occupational activities contemplated by the individual.

Perhaps the first step in vocational adjustment should be the crystallization of those interests in the individual as they relate to activities integral with a given job. Investigators in the field of measurement have long been concerned with vocational interests, but no previous attempt has been made to focus specific job-activity preferences on an occupational pattern. It is these data, after all, which point to the initial job, determine the individual's interest or boredom in his first activities, and determine to some extent his progress in it.

Interest measurement has been confined largely to the matter of general interests, and although such interests may be suggestive of occupational areas to be considered, they cannot be regarded as motivators of initial occupational activity.

There is great need today for more specific measurement which will indicate initial jobs compatible with an individual's interest in specific activities.

The construction of the *Primary Business Interests* test is direct, functional, and highly specific in an area of available business positions for beginners. The individual is asked to express his preference or dislike for those specific job-activities which are characteristic of such beginning jobs. Such activities have not been empirically selected, but they were based on an extensive analysis of beginning positions, and only those specific activities which determine and differentiate definite occupational patterns were retained.

Unfortunately no such functional occupational classifications are available — a fact which may explain why so obvious and direct an approach to interest measurement has not been used before. Actually no material changes in the matter of item selection have occurred in 20 years, and the first inventory appearing in 1921 under the name of the *Carnegie Interest Inventory* has set the pattern of which practically all general inventories now available are merely revisions. Scoring methods, it is true, have become highly refined, but it is doubtful if statistical refinement in scoring is any substitute for item-validity or basic data.

Obviously the most desirable way of selecting items for any instrument would be to set the initial research so that the items would evolve as a matter of research rather than empirical choice.

With this in mind, an extensive study was made of initially available business jobs. This study was based on a classroom assignment given to first-year evening college students who had several months' experience in business. Each student kept a work diary of a typical week and was given a grade based on the specificity of detail contained. 106 different jobs covering a considerable range of activity were analyzed, and over 2000 specific items were listed. Items which occurred less than five times in the data were eliminated. Remaining items were reduced to terse expressions of the actual activity. Reduction

PRIMARY BUSINESS INTERESTS TEST

in number of items was first achieved by grouping on a basis of a standard terminology wherever possible. Further reduction was based on the concomitance, similarity, and simultaneity of occurrences. The question of whether two items implied the same activity or invariably occurred together was decided by the majority opinion of five vocational counselors. Judgments were expressed independently on forms provided for that purpose. All counselors were actively engaged in placement work.

The list of job-items which appeared in the job-analysis form was the result of the early work in this study. The data used in its construction now have no further significance. The job analysis was used a year later with a different group of evening students who were similarly employed and similarly motivated. As the job analysis was in the form of a check-list, the data appeared in easily tabulated form. A page was attached to cover the listing of any item which may have occurred on the job, for which a printed item was not provided. However, analysis of these additions did not reveal frequencies high enough to warrant their inclusion in subsequent statistical work.

A word of explanation as to the check columns on this form may be in order, although the instructions clearly cover them. These instructions were the same as those used for the final form and are given on page 132. The first set of columns provided a quantitative analysis of items occurring on the job. The instructions given for the second set of columns, however, were designed primarily to motivate the individual. Actually these columns had a definite research purpose to fulfill. The information they provided as to how often the items occur in the job ahead was an important factor in determining their ultimate importance. Nearly 300 job analyses were received, but only 245 of these covered positions initially available in the field of business. But before proceeding with the statistical analysis of these data, let us first review what has been done within the field of occupational classifications.

We have indicated the soundness of an interest test based

on an individual's preference for specific job-activities. In order to score such a test, however, we are concerned with the way and manner in which these activities may be grouped into occupational patterns.

Occupational Classifications

The various methods of occupational classifications are largely empirical in nature. The majority are based on the census classification. This grouping is primarily concerned with economic factors. The importance of such a classification cannot be overlooked since a large part of the statistics available are based upon it.

From the point of view of similarity of activities, the census classification is of little use. To associate musicians and osteopaths, or showmen and college presidents under the same heading gives no clue as to functions involved. Under trade, to follow advertising agencies by stock-yards is similarly of little help in understanding the nature of the activities. A further weakness is illustrated by listing together inventors and draftsmen under professional service; no concern is evidenced for occupational levels.

With these difficulties in mind, an improvement on this classification was reported by Edwards¹ at a recent meeting of the American Statistical Association.

A more elaborate outline was presented at the same meeting by Palmer.²

Kimball³ redistributes the number of gainfully employed as reported in the census figures, by percentages. Reduced as it is to socio-economic groups, his study is serviceable in showing certain shifts of employment.

Similar classifications arise to expedite the very important

¹ Alba M. Edwards, "A Social-Economic Grouping of the Gainful Workers of the United States," *Journal of the American Statistical Association*, XXVIII (Oct. 26, 1940), p. 378.

² Gladys L. Palmer, "The Convertibility List of Occupations and the Problems of Developing It," *Journal of the American Statistical Association*, XXXIV (1939), p. 700.

³ Bradford F. Kimball, *Changes in the Occupational Pattern of New York State*. (Albany, New York State Education Department. Educational Research Studies No. 2, 1937), p. 38.

PRIMARY BUSINESS INTERESTS TEST

matter of recording and tabulating job placements. For example, the Massachusetts State Employment Service classifies their placements by industrial groups, very similar to the census classifications, as well as by occupational groups.

The *Dictionary of Occupational Titles* divides the major occupational groups into seven classifications, arranged alphabetically and identified by the first and second digits of the code numbers. Job classifications within these major groups are identified by three digit groups following the first two code numbers.

Humphreys⁴ gives us another classification which is concerned with the general functional aspects of jobs as they apply to many industrial and commercial establishments. This classification groups functional activities regardless of the industry or field in which they are found.

The counselor is more concerned with a similarity of the clerical functions in different fields than with the differences between the fields themselves. It is the actual function of the job which is significant to the prospective worker. This concern with the worker leads to other methods of classifications. Kelley⁵, Thurstone⁶ and others would classify occupations by the pattern of abilities required. This concern with multiple abilities will have far-reaching effects in occupational choices on the basis of profile-matching if the vocational application of such profiles is ever understood.

Kitson⁷ attempts to group vocations by the kind of training required, and it is to be regretted that this approach has not been followed up by more specific work based on classifications of pre-entry requirements quantitatively expressed

Brewer⁸ gives us a three-dimensional concept of occupations classified by fields, functions, and occupational levels.

⁴ J. A. Humphreys, *How to Choose a Career* (Chicago, Science Research Associates, 1940), 48 pp.

⁵ Truman L. Kelley, *Essential Traits of Mental Life* (Cambridge, Harvard University Press, 1935).

⁶ L. L. Thurstone, "A Multiple Factor Study of Vocational Interests," *Personnel Journal*, No. 10 (Oct. 1931), 198-205.

⁷ Harry Dexter Kitson, *The Psychology of Vocational Adjustment*. (Philadelphia, J. B. Lippincott Co., 1925).

⁸ John M. Brewer, *Occupations* (Boston, Ginn & Company, 1936), 437-441; 590-597

This limited treatment of the various approaches serves to illustrate a variety of classifications including industries, socio-economic factors, intelligence, abilities, and general interests. From the guidance point of view we are not concerned with the socio-economic status, as the number of trained workers in each occupation tells us little of its function. Classifications on the basis of abilities and intelligence give us more definitely an idea of the requirements of occupations, but disregard underlying interest in such activities. Classifications based on broad interest patterns fail too in that such patterns express only a broad attitude rather than immediate and specific interests in the actual work of the beginning occupation.

A method of occupational classification of inestimable value would be one which brought together those jobs which call for the same specific activities in approximately the same proportion, regardless of field or title. In only a very general sense can we assume that either a mention of the field or occupational title gives any indication as to what is inherent in the job itself. In fact, we may regard such descriptions as often adding confusion to an already complicated picture. The only sound basis of grouping these occupations lies in the specific nature of the work, and accordingly calls for the same general structure of interest, skills, and personality traits on the part of the worker. Although the psychometrist has accomplished much from a diagnosis of an individual's personal qualities, he has, as yet, been unable to indicate the social or economic significance of these same qualities. In the last analysis it is the latter phase which makes the first meaningful.

This study presents a statistical approach to such a functional classification but within a limited range of occupational activity. Its purpose is to measure the relationships between specific job-activities of initially available business positions and to discover what common factors exist so that special functions may be isolated. The activities which in themselves are closely related and conversely alienated from the others form the patterns needed as a scoring scheme for our interest test.

The same method of pattern determination is equally ap-

plicable to other threshold positions, as well as various occupational levels. The extension of this work would be infinitely worth while and of far-reaching consequence in the field of guidance.

Setting Up a Contingency Table

Our data consist of 245 analyses of initially available business jobs. The specific activities have been checked in each analysis in such a way as to indicate whether the activity occurs *much* or *occasionally* in the job. Before a computation of the interrelationships of these items can be made, a tabulation of the number of times that each item occurs in the data as well as the number of times which each item occurs with every other item must be recorded. That is to say, we must know how often item No. 1 occurs, also how often it occurs with 2, 3, 4, 5, and up to 115. We must know how many times item No. 2 occurs with 3, 4, 5, etc., and similarly until we have a table of all such contingencies.

A contingency table so constructed gives us at a glance a frequency of concomitance of any item with all others. Since this particular table comprises 6,550 cells above the diagonal, space hardly permits its inclusion here. The diagonal values themselves represent the number of times that each item occurs in the 245 analyses and the largest value that any can take is, of course, 245.

In constructing this table the I.B.M. tabulating equipment was used, a Hollerith card being punched for each case.

The card provides 80 columns which may be punched from zero to nine in each, and above this range of digits two other positions may be punched, making in all twelve positions within a column. In this instance, the first three columns were used to carry the number of the job-analysis form. For example, form 133 has 1 punched in the first column, 3 in the second, and 3 in the third. In the fourth column the items between 1 and 9 occurring as either *much* or *occasionally* on the job-analysis were punched; in the fifth column number 10 was punched as zero, 11 as 1, up to 19 as 9; in the sixth column

items 20 to 29; and similarly for the remaining. The top two positions were not used since dichotomies of ten items in each column facilitated rapid punching and eliminated more complex conversions. After cards were punched, they were checked by another clerk. The resultant contingencies formed the basis for succeeding statistical work.

Weighting the Job-Items

Our contingency table, comprising some 6500 cells, gives us graphically the raw count of the occurrence of each item with every other. We cannot assume, however, that all items are of equal importance. Before going further with the not inconsiderable computational work of pattern-determination, let us see what items might now be eliminated as of little importance in resultant groupings. What factors are significant in making this decision?

Obviously, the frequency of occurrence of the item, indicated as a diagonal value, must be considered; so, too, its variance is a natural weighting factor. Other *à priori* considerations also occur which may or may not be inherent in the data. Of a list submitted to a group of placement officers these four were considered most important:

1. Proportionate time devoted to each item on the job.
2. Need in job-activity of position ahead.
3. Amount of training required previous to employment.
4. Relative importance in the selection of the employee

The first two of these considerations could be determined from the data, since the job-analysis form used and previously described provided check columns, so that an *M*, *O* or *R* (much, occasionally, or rarely) response could be recorded in respect to each of these considerations. The third and fourth considerations, however, could not be objectively determined. Five vocational experts were, therefore, asked to rate each job item on a three-point scale in these respects. The following paragraphs are quoted from the written instructions to the

PRIMARY BUSINESS INTERESTS TEST

judges, which were further clarified in a group meeting before the ratings were made.

"Amount of training involved" The phrase refers to the amount of training received before employment and considered necessary by the employer to perform the job activities involved. Such training naturally differs in amount, which may be illustrated by such items as "post book-keeping entries," in which case the employer would expect the individual to have had some training, and such an item as "make up balance sheet" in which case considerable training previous to employment would have been necessary. Please consider carefully, therefore, each of the 115 items in respect to this consideration, using the first column for your check mark if you consider a relatively large amount of training is involved previous to employment and using the second column if you consider a lesser amount involved. Make no check marks if the amount of training previously necessary is relatively little.

"Relative importance in selection of employee." The ability to do certain of these items may be an important consideration in the selection of an initial employee. This is particularly true in respect to job-activities of the contact type. Make no check marks if the ability to do the activity is of relatively little importance in selection. Use the fifth column where ability to do this item becomes of importance in selection and step up your check mark to the fourth position if you feel that it is of considerable importance in employee selection. To illustrate, the question of being able to "call on clients" is undoubtedly a factor in selection. But an item such as "sell goods on commission basis" is considerably more important, the ability to do which is probably the primary consideration in the selection of an employee where such an item probably constitutes his chief activity."

The weighting formula for each item is composed of these six considerations. Since each cell is determined by the concomitant occurrence of two items, it consequently has a weight equal to the product of the weights of the two diagonals which compose it, and no such result, of course, can exceed the product of the square roots of the weighted diagonals; can in fact equal it only when there is maximal concomitance of any two items in our data. The formula for the Aggregate Weight⁹ of each cell value is, therefore, composed of the square roots of each frequency, the square root of each variance which in the case of a proportion is pq , and an additive weight of factors just considered for each of the contingent items.

⁹ The vital assistance of Professor T. L. Kelley of Harvard in developing the statistical procedures of this study is gratefully acknowledged.

$$(5) \text{ Agg. Weight}_{12} = (W_1 \sqrt{f_1 p_1 q_1}) (W_2 \sqrt{f_2 p_2 q_2}) \quad (1)$$

in which $W_1 = w_a + w_b + w_c + w_d$ and

$$w_a = \frac{M}{M + O}, \quad (2)$$

$$w_b = \frac{2(pu + su) + (=)}{2(f + pu)}, \quad (3)$$

$$w_c = \frac{2Lt + St}{2j}, \quad (4)$$

$$w_d = \frac{2Cs + Ss}{2j} \quad (5)$$

Formulae 2 through 5 take on values between 1 and 0 and are additive in their function. These same numbers will identify each formula with the consideration afore-described. The symbols are thus interpreted. In (1) M is the number of times the item occurs *much* on the job, and O *occasionally*.

In (2) pu refers to the *pick-ups*, or frequencies of occurrence in job ahead but not in present job, while su indicates *step-ups* in the amount of the activity in the next job over the present, and $=$ refers to the tabulations of equal amount of activity in job ahead as in present job.

In (3) Lt indicates the number of judgments that a large amount of pre-entry training is required, St a somewhat smaller amount; j the number of judges.

In (4) Cs indicates the judgments that ability to perform the job activity is a considerable factor in selection of employee, and Ss somewhat of a factor in selection.

These aggregate weights range as low as .44 on item 105 to as high as 8.07. Items 12, 15, 16, 74, 78, 79, 105, 106, 108, and 112 were eliminated because of low weights as definitely of little importance in further consideration. These ten items had weights below unity and/or frequency of 12.

Computing Correlation Coefficients

In order to determine the relationships that existed between these contingent frequencies, the cell values given in

PRIMARY BUSINESS INTERESTS TEST

the contingency table were converted into the more usual correlation mold. It may be observed that the raw tabulations in the contingency table can be expressed as p 's, proportions, computed by dividing the observed values by 245, which is the total number of times it could have occurred in our sample. This table of p 's can next be converted into co-variances by subtracting from the cell p the product of the diagonal p 's occurring in its row and column. This p_{12} minus p_1p_2 becomes the numerator of the now easily recognizable formula for the product-moment correlation, the denominator being the standard deviation of each variable, which in this special case is \sqrt{pq} of each. These values are readily obtained from the new Kelley tables¹⁰ for any given p . For the contingency method, then, of computing correlation coefficient we have

$$r_{12} = \frac{p_{12} - p_1p_2}{\sqrt{p_1q_1} \sqrt{p_2q_2}}, \quad (6)$$

$$\text{in which } p_{12} = \frac{f_{12}}{N} \text{ (cell)}$$

$$p_1 = \frac{f_1}{N} \text{ (diagonal)}$$

$$q_1 = 1 - p_1$$

$$N = 245$$

The resulting intercorrelation table gives all positive correlation values of .3 or better, and negative correlations of .2 or greater. The dots occurring in other cells indicate smaller values and fall between —.20 and +.30. All values below this diagonal, as in the contingency table, are omitted in the interest of economy, merely duplicating as they do the values above the diagonal.

It will be observed that in only one instance does any value equal .80. This fact would seem to indicate that the work of combining job-items on the basis of concomitance and simultaneity was adequately done. It may be said that a correlation

¹⁰ *The Kelley Statistical Tables*, (New York, The Macmillan Company, 1938).

closely approximating unity would be of little value in pattern determination, since it would indicate invariably that one job-activity depended entirely on the other or always occurred with it.

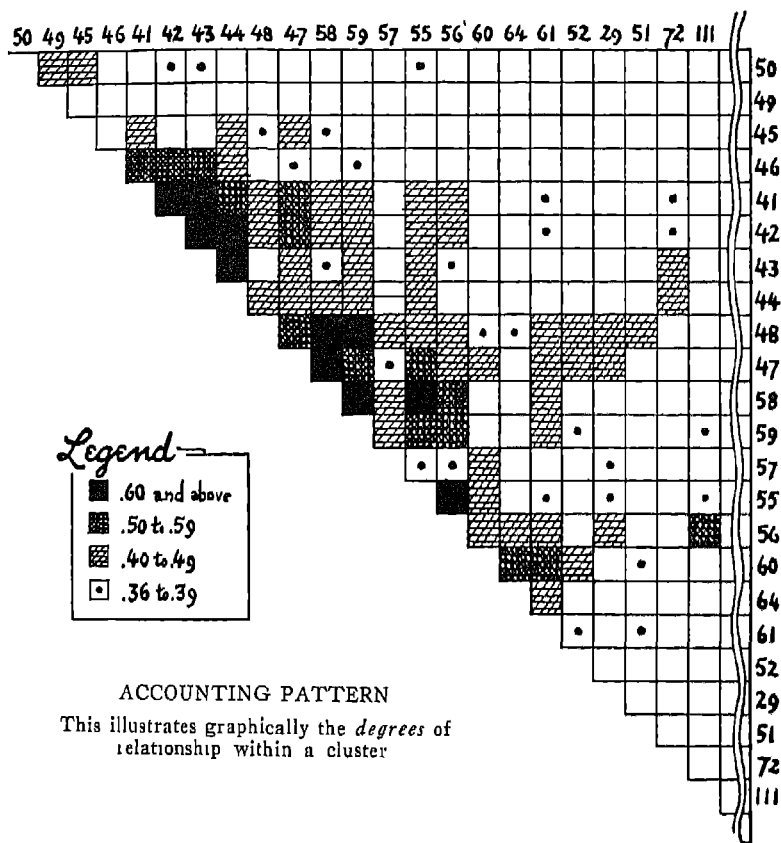
Grouping Correlation Coefficients

Our task now becomes one of discovering what functional patterns of occupational activities exist in our data. The problem of a cluster analysis may be visualized as a plateau of relationships which, if rearranged in rows and columns, would result in a topographical map; high relationships forming a peak surrounded by related and interrelated activities. In a sand-pile graph, low relationships form the valleys and do not help in distinguishing one cluster from another. In short, these correlation coefficients must be so inter-changed within the matrix that the highest value in each row and column appears as near the diagonal as possible. If such patterns as we have postulated are inherent in the data, this re-arrangement should result in clusters along the diagonal of the new matrix.

A table representing this re-arrangement was constructed showing that several clear-cut clusters or patterns are now evidenced. Values of .40 and above occurring in a row and column appear to indicate the extent to which an item "belongs" to a pattern, and very few such values occur very far removed from the diagonal. A closer grouping of these items, however, is possible by the further elimination of such items where no value of .40 or better occurs in its row and column, and where two items, such as 107 and 109, while closely related, appear to be nearly discrete in themselves. Twenty-nine items may now be eliminated as having no value as great as .40, leaving only 75 items which fall into specific groupings. In a very few cases items have values too high to be discarded, but appear to belong equally well to two patterns. In delineating these patterns, therefore, these items will be given one-half weight in respect to each pattern.

PRIMARY BUSINESS INTERESTS TEST

This table is too extensive to reproduce here but a single pattern will give a graphic illustration. By converting figures into toned equivalents, the size of a correlation coefficient becomes a density value which illustrates the degree of relationship within patterns.



Following this chart are tables in which each cluster has been treated as a single matrix, so that the relationship between the items within it may be more carefully studied. The name of each pattern has been determined by the most apparent function within it.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 1

ACCOUNTING

49	45	46	41	42	43	44	48	47	58	59	57	55	56	60	64	61	52	29	51	72	111
42																					
			42			40		40													
	50	55	51	43																	
		80	64	56	45	53	44	46			44	42									
			71	65	45	53	46	48			46	40									
				69		46		47			41									46	
					48	47	44	48			40									42	
						56	63	65	44	46	48					45	42	46	42		
							61	59		52	49	44					47	40			
								79	48	63	59				41						
									47	58	57										
													46								
													66	45							
													46	45							
														45	41			44		50	
														50	50	48					
															43						

- 50 Verify bookkeeping records, audit, etc.
- 49 Set up new system of accounts
- 45 Make up tax returns
- 46 Keep inventory records
- 41 Make bookkeeping entries
- 42 Post bookkeeping entries
- 43 Take off trial balances
- 44 Make up balance sheet profit and loss statement
- 48 Make out and figure payrolls
- 47 Make out monthly statements, bills, figure extensions
- 58 Reconcile check book with bank statement
- 59 Enter checks received in check book
- 57 Take care of petty cash
- 55 Draw checks
- 56 Make out deposits
- 60 Figure trade discounts, commissions, etc.
- 64 Figure salesmen's commissions
- 61 Figure interest
- 52 Check invoices, prices, discounts and allowances, etc.
- 29 Determine credit risks
- 51 Prepare special reports, sales analyses, etc.
- 72 Type financial statements
- 111 Take deposits to banks, cash checks, collect bank statements and have checks certified

PRIMARY BUSINESS INTERESTS TEST

TABLES 2 AND 3

COLLECTIONS AND ADJUSTMENTS

26	25	27	28	68	
	50				14
	43				26
		36			25
			56		27
				40	28
					68
14	Call on clients				
26	Make personal calls for credit information				
25	Make personal collection calls				
27	Make customer adjustments, smooth out difficulties				
28	Handle complaints, investigate				
68	Telephone delinquent customers				

JUNIOR CLERICAL

69	111	98	99	103	102	101	104	84	100	
52						46				86
			41							69
			51	42						111
			54				40			98
				65	45		44			99
					58	45	51		45	103
						59			42	102
										101
								46	40	104
										84
										100
86	Mail out statements and correspondence									
69	Address envelopes, bills, etc.									
111	Take deposits to banks, cash checks, etc.									
98	Pay bills and bring back receipts									
99	Get mail and stamps at post office									
103	Take mail, registered mail, and parcels to post office									
102	Seal, weigh, and stamp mail									
101	Fold letters, circulars for mailing									
104	Run errands for employer									
84	Assist others, general handyman									
100	Take messages, papers, and distribute mail									

TABLES 4 AND 5

STENOGRAPHIC — FILING

86	69	71	67	65	66	96	87	72	
	52								86
		51	40						69
				60		45		49	71
					43				67
									65
									66
									96
							46		87
									72
86	Mail out statements and correspondence								
69	Address envelopes, bills, etc.								
71	Type letters, orders, forms, etc.								
67	Answer telephone								
65	Take dictation and transcribe								
66	Code and type telegrams								
96	File orders, letters, bills, reports, trade information								
87	Look up information in files, library, etc.								
72	Type financial statements								

SALES — OFFICE

22	24	62	90	19	91	13	93	97	94	92	
	41										22
		40									24
			42								62
				37							90
					40						19
						56					91
							84	77	53	47	13
											93
											97
											94
											92
22	Make out price sheets										
24	Check on competitors' prices, compare quotations										
62	Figure quotations										
90	Dictate letters, reports, etc.										
19	Attend conference with supervisor										
91	Organize work, train and supervise others										
13	Make up sales contracts										
93	Classify orders to size, patterns, salesmen, etc.										
97	Keep trade information, credit information, up to date										
94	Make forms and charts										
92	Purchase merchandise, supplies, equipment										

PRIMARY BUSINESS INTERESTS TEST

TABLE 6

SALES — STORE

7	33	37	4	114	5	40	11	31	1	2	20	9	8	6	113	
																38
																10
																39
																34
51																7
40																33
	45															37
																4
																114
																5
																40
																11
																31
																1
																2
																20
																9
																8
																6
																113

- 38 Make up and schedule shipments, etc.
- 10 Deliver orders to customers
- 39 Put up mail and telephone orders, fill orders to be shipped
- 34 Check on quality of goods, examine for defects
- 7 Take inventories
- 33 Check and receive incoming supplies, record, etc.
- 37 Unpack goods, put away and keep storeroom in order
- 4 Wrap up bundles and packages
- 114 Sweep floors, empty waste baskets, clean up, etc.
- 5 Put tags and labels on merchandise
- 40 Dust shelves, put in order
- 11 Restock shelves and cases
- 31 Give information, quote rates over telephone
- 1 Wait on customers, sell over the counter
- 2 Sell goods over telephone
- 20 Letter signs for stock display
- 9 Set up displays, window trim, etc.
- 8 Dismantle window displays
- 6 Arrange display of food stuffs
- 113 Clean refrigerator, show cases, equipment

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Final Occupational Patterns

The foregoing tables comprise the resultant six occupational patterns, namely, *accounting, collections and adjustments, junior clerical, sales-office, sales-store, and stenographic-filing*. These correlation coefficients, however, indicate only the "belongingness" of items within their pattern. The relative significance of each item within the pattern is determined by returning again to the table of weights. We now relist these items under their proper pattern headings with their respective weights as determined earlier in this study.

TABLE 7

ACCOUNTING		Weight
3	Verify bookkeeping records, audit, etc.	3.11
5	Determine credit risks	2.97
12	Make up tax returns	2.59
16	Make out monthly statements, bills, figure extensions	4.21
21	Make up balance sheet, profit and loss statement	3.08
23	Make out and figure payrolls	2.60
26	Set up new system of accounts	1.25
27	Figure interest	2.23
29	Make out deposits	2.42
31	Check invoices, prices, discounts, allowances, etc.	4.05
33	Post bookkeeping entries	3.83
35	Prepare special reports, sales analyses, etc.	4.28
37	Enter checks received in check book	1.31
52	Reconcile check book with bank statements	2.37
55	Keep inventory records	4.46
57	Take off trial balances	4.86
63	Make bookkeeping entries	4.63
68	Figure trade discounts, commissions, etc.	3.14
69	Draw checks	2.61
70	Take care of petty cash	3.57
74	Figure salesmen's commissions	1.82
Score $\frac{1}{2}$ on		
38	Type financial statements	2.14
6	Take deposits to banks, cash checks, collect bank statements and have checks certified	4.30
COLLECTIONS AND ADJUSTMENTS		Weight
7	Handle complaints, investigate	7.38
11	Make customer adjustments, smooth out difficulties	8.04
14	Call on clients	3.33
17	Telephone delinquent customers	3.14
20	Make personal calls for credit information	1.83
64	Make personal collection calls	2.99
JUNIOR CLERICAL		Weight
9	Pay bills and bring back receipts	3.07
36	Get mail and stamps at post office	2.74
40	Assist others, general handyman	2.94
41	Take mail, registered mail and parcels to post office	2.46
44	Fold letters, circulars for mailing	1.74
49	Seal, weigh, and stamp mail	2.68
50	Run errands for employer	2.51

PRIMARY BUSINESS INTERESTS TEST

61. Take messages, papers, and distribute mail to departments	1.94
Score $\frac{1}{2}$ on:	
42. Mail out statements and correspondence.	4.01
60. Address envelopes, bills, etc.	2.07
66. Take deposits to banks, cash checks, collect bank statements, and have checks certified	4.30
SALES—OFFICE	Weight
4. Check on competitors' prices, compare quotations.	2.66
15. Dictate letters, reports, etc.	2.25
22. Attend conference with supervisor.	3.03
39. Make forms and charts.	3.65
45. Purchase merchandise, supplies, equipment	5.91
48. Make out price sheets.	2.69
54. Organize work, train, and supervise others.	4.31
58. Make up sales contracts.	2.40
71. Keep trade information, credit information, up to date.	2.91
73. Classify orders to size, patterns, salesmen, etc.	1.45
75. Figure quotations	2.06
SALES—STORE	Weight
1. Wrap up bundles and packages.	5.50
2. Sell goods over telephone	6.10
6. Unpack goods, put away, and keep storeroom in order	4.58
8. Give information, quote rates over telephone	7.77
10. Put tags and labels on merchandise.	3.09
13. Deliver orders to customers.	2.50
18. Dust shelves, put in order	2.60
24. Restock shelves and cases	4.11
25. Set up displays, window trim, etc.	4.28
32. Make up and schedule shipments, etc.	3.98
34. Check and receive incoming supplies, record, etc.	8.07
43. Arrange display of food stuffs.	2.61
46. Put up mail and telephone orders, fill orders to be shipped	4.17
47. Letter signs for stock display	1.38
53. Check on quality of goods, examine for defects	7.53
59. Sweep floors, empty waste baskets, clean up, etc.	2.04
62. Dismantle window displays.	1.49
67. Take inventories	6.26
72. Wait on customers, sell over the counter.	7.95
STENOGRAPHIC-FILING	Weight
19. Answer telephone	5.08
28. Take dictation and transcribe.	4.92
30. Code and type telegrams	1.51
31. Type letters, orders, forms, etc.	3.95
56. File orders, letters, bills, reports, trade information	4.67
65. Look up information files, library, etc.	7.46
Score $\frac{1}{2}$ on:	
38. Type financial statements	2.14
42. Mail out statements and correspondence.	4.01
60. Address envelopes, bills, etc.	2.07

It may be observed that these weights run from 1.25 to 8.07 with the greater part of them falling between values of 2. and 5. Since only one value exceeded 7.99, unit weights of from 1 to 7 were used (determined by value without regard to decimal) in scoring each item in respect to the pattern in which it belonged.

Final Interest Test

We have now arrived at a list of 75 specific job activities which are common to initially available business jobs and which fall within specific patterns; we have determined, too, their relative significance within these patterns. An individual's reaction, therefore, as to the extent to which he feels that he would like or dislike these activities, can now be evaluated in terms of specific beginning business positions. Before setting up these items in terms of responses, however, we must consider how they shall be listed. To merely list them in the order in which they appear under their pattern headings would obviously condition responses which seemed to go together. A random order seems desirable. Accordingly the listing by patterns is now numbered from 1 to 75, and these numbers converted into a random order by Fisher and Yates' Table of Random Numbers.¹¹

On the final form¹² the questions appear directly on an I.B.M. answer sheet. Instructions are also printed on this sheet and space provided for name and pertinent information concerning the individual taking the test. The instructions are reproduced here but the items have appeared earlier, though in different order.

Instructions

This questionnaire is designed to indicate how you feel about those specific job-activities which characterize initial business positions. You are to indicate your answer by blackening the space between the proper pair of dotted lines.

The first three columns are headed L I D so that you may record your response as *like*, *indifferent*, or *dislike*. If you think that you would like to perform the job-activity indicated as a part of your first business job, record your

¹¹ R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, (London, Oliver and Boyd, 1938).

¹² Published by Science Research Associates.

PRIMARY BUSINESS INTERESTS TEST

response under L, if you feel uncertain or indifferent, under I, if you feel that you would dislike this activity, under D. Omit no items.

When you have completed this, go over the items again carefully and indicate in the X column the *five* which you would *most like to do*. There is no time limit, but you should work fairly rapidly as it is your first impression which is important.

Scoring

The L I D responses are, of course, familiar to all. Here the "L" response is scored with the weight previously determined within its respective pattern. The "I" response is scored as one-half that amount with all fractions dropped, with the "D" response scored as zero. The "X" response is new to previous practice in this type of test, and allows an opportunity for an individual to distinguish a little more carefully between those job-activities which he feels he would like as part of his initial job. It serves, too, as an additional aid to the counselor, apart from the resultant pattern scores. The fact that the individual selects five out of 75 items as *most* to be desired should result in additional weights in respect to such items. We may logically expect these selections to be made in respect to items in which an "L" response has been recorded, and an additional weight equal to one-half the weight of the "L" response should result in some refinement in scoring without being regarded as excessive. This "X" response can be scored without additional runs through the machine.

The test may also be scored by hand. Special hand-scoring folders are provided which hold the sheet in position for the proper registration. The norms appear directly on this folder which is so cut that the score has to be recorded in the right place each time. For machine-scoring two keys are provided for each pattern. By picking up a number of "contrasts" each time only two runs are necessary to pick up positive weights of from one to seven; thus machine and hand-scored norms

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

are identical. Since it is the same sheet in each instance, test conditions are also identical.

Directions for Administering

Some students will complete this test in twelve minutes; the majority, however, finish in approximately fifteen minutes, and none takes over twenty minutes. The subjects' first run through in respect to the L I D responses is done rapidly and without hesitation, and this observation leads us to believe that there is very little doubt in their minds as to how they feel about these particular job-activities. Some hesitance, however, is seen in selecting the five job-activities most to be preferred, which is equally desirable, as it indicates a tendency to weigh them carefully.

Intercorrelations, Reliability and Validity

TABLE 8

	Acctg.	Coll. & Adj.	Jr. Cler.	Sales Office	Sales Sten.	Sten. Filing
Accounting92	— .25	.08	.22	.10	.22
Coll. & Adj.73	— .13	.27	.00	— .05
Jr. Clerical78	.01	.65	.41
Sales—Office78	.26	.07
Sales—Store77	.31
Sten.—Filing80

In Table 8 we present the intercorrelations of these patterns with their reliability coefficients as diagonal values. With the exception of the relationship between the Junior Clerical and Sales-Store patterns these coefficients are low enough to indicate satisfactory independence of these patterns. The correlation of .65 between the two mentioned, however, is too high to be disregarded. Further, such raw coefficients understate the actual relationships involved. To eliminate as much of the chance factors as possible we apply the formula¹³

$$\frac{r_{ab}}{\sqrt{r_a} \sqrt{r_b}}$$
 to correct for attenuation. The corrected value of .83 clearly indicates that these patterns do not act independently. A study of the relation of these two patterns

¹³ C Spearman, *American Journal of Psychology*, XV (1904), p. 271.

PRIMARY BUSINESS INTERESTS TEST

to the other four provides still further evidence that these two patterns should be combined.

	Acctg.	Coll & Adj	Sales Office	Sten. Filing
Junior Clerical08	— .13	.01	.41
Sales—Store10	.00	.26	.31

Combining these two patterns would also serve to step-up the reliability, as this coefficient is materially affected by the range of scores made on a test.

TABLE 9

	Acctg.	Coll. & Adj.	Sales Office	Sales Strs	Sten. Filing
Accounting92	— .25	.22	.13	.22
Coll. & Adj.73	.27	.06	— .05
Sales—Office78	.13	.07
Sales—Store86	.37
Sten.—Filing80

Table 9 represents the revised matrix on the basis of five patterns with the diagonal values representing the reliability coefficients as in the preceding table. The highest relationship now observable in this revised matrix is that of .37 between Sales-Store and Stenographic-Filing, and even this is satisfactorily low. Considered critically, it becomes .44 when corrected for attenuation; squaring it, it becomes .19, and gives us an idea of the variance which can be accounted for in this relationship. More significant in our present consideration is the amount of variance not accounted for, given by $1 - r^2$ — in this case .81.

It will be noticed that the combination of the two patterns which were not independent now has a reliability of .86. The lowest reliability coefficient occurs in the Collections and Adjustments pattern with a .73, having only six items, and the highest reliability coefficient in Accounting with .92, a pattern comprising 23 items.

As has been pointed out, the size of a reliability coefficient is partially determined by the number of items, particularly when computed by the method of split-halves¹⁴, and it might be argued that more items should have been retained in setting

¹⁴Karl J Holzinger. "Note on the Use of Spearman's Prophecy Formula for Reliability," *Journal of Educational Psychology*, XIV (1923), 301-305.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

up these tests. The retention of such items, however, would have been at the expense of the validity of the test, since the test was constructed with this objective initially in mind. This approach deviates materially from the common practice in the construction of interest test questionnaires, which are usually built from items empirically determined, and then scrutinized to determine their validity. Items, of course, are usually rated before inclusion in regard to relevance and importance, and the ultimate validity determined by the analysis of scores obtained by persons successfully engaged in such occupations, or still more feebly, by the opinion of experts as to internal consistency. The validity of these items was determined by a cluster analysis, and only those which acted as pattern-determiners were used in the test.

Immediately following are norms for these final five patterns in tabular form. They are given as standard scores and normalized scores with $M = 50$ and $\sigma = 10$, and are based on 304 freshmen at Boston University, College of Business Administration. Additional norms are available but are not necessary in the clinical use of the instrument. What is important is the "level of significance" which is here considered as the upper half of the range.

TABLE 10

Percentile	Standard Scores	Raw Scores			
		Acctg.	Coll. & Adj.	Sales—Office	Sales—Store
100					100
99	2.32			36	91
98	2.05	74		34	84
95	1.64	68	30	32	77
90	1.28	62	27	30	69
80	.84	56	23	27	62
70	.52	50	21	25	56
60	.25	46	19	23	51
50	.00	42	17	22	46
40	— .25	38	14	21	41
30	— .52	34	12	19	35
20	— .84	28	10	17	29
10	—1.28	22	6	14	22
5	—1.64	16	3	12	14
2	—2.05	10		9	7
1	—2.32	6		7	2

PRIMARY BUSINESS INTERESTS TEST

Present Use of This Instrument

This test is now being used for several purposes. First, it assists the freshmen of the College of Business Administration in the selection of academic majors. Interest in the particular job-activities involved in the occupational areas for which such majors train is by far the most important single consideration in making such selection. True, the probable "risk" of pursuing such training in terms of special abilities and personality traits still remains to be evaluated. Many instruments, however, are available for diagnostic purposes, and each entering freshman now regularly takes a battery of such tests upon admission.

Second, it was used in the Evening Division of the College of Business Administration of Boston University where the immediate pursuit of an initially available job is economically imperative. The "X" response in particular has been found helpful to the counselor when the student comes to his desk, and the whole approach seems to better motivate the individual's interest in the specific job-activities of beginning positions which he has been considering tentatively.

For commercial students at the senior high school level who are looking for jobs the results should have the same significance as for the evening student just mentioned; they, too, need indicators for beginning positions.

The test is also being used experimentally at the 9th grade level within the commercial curriculum to distinguish between bookkeeping, sales, and stenographic interests where such help is badly needed. In situations such as found throughout New York State where a course called "Introduction to Business" is given to all commercial ninth-graders, no problem as to terminology of test items is met; but where no such orientation course is given, several terms need discussion and explanation. The maximum usefulness of this direct testing technique rests on an extension of realistic work experience provided for in the curriculum.

In out-of-school situations this test likewise has several applications. Social agencies, faced with a need for specific

counseling previous to job hunting, find it extremely useful as a direct pointer towards specific beginning jobs. In the Guidance Department at the Boston YMCA, for example, where an excellent and extensive job in counseling is being done, this test is basic to all batteries. All members of Darling's "Job Hunters" group at Boston take this test since a large part of them are interested in beginning business jobs. Employment managers likewise in several situations are using this test for beginning office workers where the present scarcity of beginners places more emphasis on allocation than merely selection.

Implications Arising From This Study

The usefulness of this test is obviously limited within the range of an already generally determined interest in the field of business. What is still needed is a general interest test of this functional type which would allocate student interest into general areas. This accomplished, areas needing a more detailed diagnosis would be indicated. An extension of the horizontal testing range for initially available positions is equally feasible in other areas. As has been previously indicated, maximum effectiveness of this positive approach to interest measurement depends upon the progressive development of the secondary school curriculum toward increased emphasis on realistic bits of work experience. Progressive educators predict that within a few years 25 per cent of the ninth-grade work may be of this type, increasing until 75 per cent of the twelfth-grade level curriculum will be composed of such work experiences. Such a program will go far toward bridging the present gap between school and job placement.

It should be similarly noted that a vertical extension of this testing technique is equally possible. Actual experience on a job results in a refinement of interests in the field and should be measurable as a pattern indicative of progressive specificity. Many quite different jobs may be indicated from the development of more specific interests within a single initial pattern, and assistance in the selection of secondary level jobs could thus be provided. The technique for constructing such interest tests is now available.

MEASUREMENT IN RURAL HOUSING

A PRELIMINARY REPORT¹

CHARLES I. MOSIER
Social Security Board

THE SLOAN PROJECT in Housing Education at the University of Florida is an attempt to investigate the broad problem of the extent to which educational materials introduced through the school may exert an influence on the social, cultural and economic life of the community as a whole. Specifically, the problem investigated is the effect on the housing status of the community, and on the housing attitudes of its members, of a broad program of education in all aspects of housing introduced through the schools. The general outlines of the program involved the utilization of six white school communities of Florida, three experimental and three control. The plan of the experiment involves a determination of the present status of all six communities in those attributes which might be affected by housing education, the introduction of housing material into the regular curriculum in the schools of three experimental communities, and subsequent tests to determine the change in housing status, attitudes, and school achievement which can be attributed directly to the experimental program.

The present report is concerned only with a summary of the initial measurement aspects of the experiment.² All measurements have been made in both experimental and control communities at the beginning of the experimental program to

¹ The study reported here is being carried on at the University of Florida under the auspices of the Sloan Foundation in Applied Economics.

² A full report of the initial measurement program has been prepared and is on file at the University of Florida Library. C. I. Mosier, *Measurement in the Field of Rural Housing* (1941).

establish a base line from which to measure change, and they will be repeated at subsequent intervals throughout the course of the experiment as well as at its close. Certain of the measurements might well be made again several years after the termination of the experimental work of education in order to determine the stability of the results and to provide a measurement of those changes occurring after relatively long latent periods. As an instance of the latter, we might hypothesize that one of the outcomes of the experiment would be that those pupils who had been exposed to the experimental program of housing education would, on leaving school and establishing homes of their own, secure more adequate housing than those who had not. The observation of the full impact of such effects could take place only twelve to fifteen years after the beginning of the experiment. Twelve years would be required for the pupil to progress normally from the first grade to high-school graduation and thus receive the fullest benefit of the educational program. At least three more years would elapse before any sizable proportion of the subjects would have married and become sufficiently established economically to provide themselves with homes which could be considered indicative of their ultimate housing status. It is not proposed that such an extended period of observation is essential to the program, but attention should be called to the long-range character of certain of its effects.

The measurements which have been undertaken can be divided into four major areas: housing adequacy, housing attitudes and insight, housing information of pupils, and academic achievement. For the measurement of housing adequacy we have developed a *Housing Inventory*³ describing the objective condition of the house as observed by a trained field-worker, yielding a *Housing Index*—a composite score for the house obtained by weighting and combining these observations to provide for each house a single score. The inventory records

³ For a detailed report on the development of the *Housing Inventory*, see C. I. Mosier, *Measurement of Rural Housing Status*, in preparation.

MEASUREMENT IN RURAL HOUSING

obtained by interview have been supplemented, in a large proportion of the cases, by photographic records of the houses studied.

It is conceivable that the program of education might produce a real change in attitude toward housing, in insight into the present inadequacies where they exist,⁴ and in motivation toward better housing conditions, and yet there might be no externally observable improvement because of the pressure of economic circumstances. Because of this, an attempt has been made to measure the extent of such effects by a separate evaluation of the answers to certain of the inventory questions. Plans have been made for a more direct measurement of attitudes and for the development of tests measuring achievement in the acquisition of information in the field of housing, but these plans have not yet been fulfilled.

It is important to know whether the introduction of housing materials into the curriculum has been at the expense of training in the fundamental academic achievements, or whether it has resulted in more effective learning of the skills of reading, arithmetic and language through a use of material of more immediate interest and appeal than that contained in the customary curriculum. As the initial phase in the investigation of this problem, a program consisting of four achievement tests and an "intelligence" test has been administered in grades 4-12 in all schools. The results of this initial testing and the subsequent retesting which is contemplated will provide valuable information on this question and on others which may arise.

Before beginning a summary of the initial results in each of the four areas of measurement, certain further general statements should be made concerning the communities investigated and the nature of the sample involved. While the detailed description of the several communities can best be presented in the light of the results of the initial surveys, cer-

⁴In answer to the Inventory question, "What changes or improvements do you think are needed", the occupant of one extremely dilapidated shanty said, "Well, if I could get hold of some cardboard boxes to tack up inside to cover the chinks, I reckon I'd have a right tight little place."

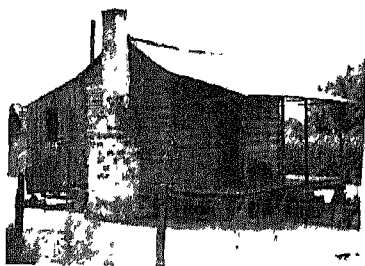
tain broad generalizations can be made which will facilitate the understanding of these results. For the purposes of this study, a community is defined as consisting of all those families, and only those families, which send at least one child to a specific school being studied. This automatically restricts the population to white families. Such a definition necessarily involves a somewhat different use of the term "community" from that ordinarily envisaged. It excludes all those families, maintaining their identity as family units, which have been established so recently that there is no child of school age. It excludes any families who have children of school age, but who, for one reason or another, do not send their children to the school in question. In one of the communities the field-workers reported this situation: "Are you going to see the Joneses down the road? They've got a flock of kids, but the kids don't go to school because they ain't got no clothes." The extent to which this, and other comparable situations, prevail is, at present at least, unknown, but it does exist and influences the sample studied, since the group under discussion will consist of the poorest families in the area. In certain of the communities the definition of the population imposes another restriction in the outlying districts. As the distance from school becomes great, the children of grade-school age go to the local school, so that only those families with at least one child of high-school age are included. The inclusion of a family in the "community" depends, then, not only on the age-distribution of the children (and hence the length of time the family has been established), but on geographical location as well. These selective factors will, inevitably, limit the extent to which the results of this study can be generalized to the community-as-a-whole, since the population studied is limited to those families sending at least one child to a participating school.

The unit of the investigation, when it is not the individual pupil in the school, is the dwelling group. All persons living within the same dwelling-unit (house or separate apartment) are considered to constitute a "family unit." A total of 745

HOUSES AT SELECTED LEVELS OF HOUSING INDEX VALUE



Score 17



Score 20



Score 23



Score 26



Score 29

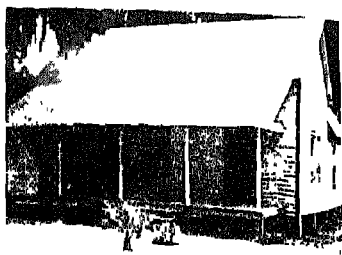


Score 32

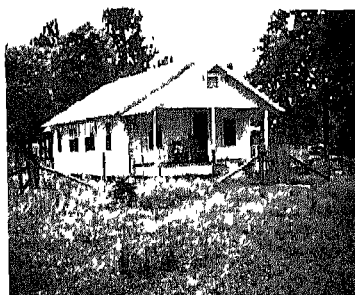
HOUSES AT SELECTED LEVELS OF HOUSING INDEX VALUE



Score 35



Score 38



Score 41



Score 44



Score 47



Score 50

MEASUREMENT IN RURAL HOUSING

families was studied, with test records of 1028 children of school-grade above the fourth. There are records of 522 children in grades 1-3, but these records are incomplete and do not represent the total number of children in those grades. The population was defined as of the date of interviewing (October and November, 1940), and families moving into the community after that date were not considered.

All of the primary data and much of the derived data have been recorded on electric accounting equipment cards, where they are readily available for further research.

The evaluation of the housing status of the communities was by means of a *Housing Inventory* specifically designed to measure housing adequacy in rural areas. In addition to identification data the *Inventory* recorded the responses to 85 items of the type:

Fireplace and chimney—state of repair? 1. no fireplace. 2. poor—masonry cracked, mortar crumbling, many loose bricks. 3. fair—masonry discolored, occasional loose bricks. 4. good—no obvious repairs needed.

Kind of cookstove (if more than one, mark the highest number). 1. open fireplace. 2. makeshift stove—sheet-tin arrangements, etc. 3. small wood stove. 4. large wood range. 5. electric. 6. gas, bottled or city.

"How many rooms, not counting closets, porches (or the bathroom) are there?" (check answer with your own observation). 1. one. 2. two. 3. three. 4. four. 5. five. 6. six. 7. seven. 8. eight. 9. nine. 10. ten. 11. eleven. 12. twelve or more—specify.

With the exception of answers to certain specified questions the observations recorded were those of a trained interviewer, not of the occupant.

The criterion which such an instrument should fulfill was established, available sources searched for suggestions as to possible items, and a preliminary edition prepared. This preliminary edition was subjected to the editorial scrutiny of research workers in the field of rural welfare, revised, and subjected to a searching field-test in which two *Inventories* were completed for each house under conditions comparable to those of the main study. The *Inventory* was revised again in the light of the field experience, and prepared in final form.

Three interviewers meeting a set of pre-established qualifications were used. They were given intensive training in the use of the *Inventory* under actual field conditions and in establishing common standards of evaluation. A booklet of *Instructions to Raters* was prepared for their further guidance. All coding of the data and computations were performed separately from the interviewing, and a routine of computation established. An occupational code adapted to the needs of the particular problem has been devised, and applied to the classification of the heads of the families studied.

The data from the *Inventory* have been recorded on punch cards. A procedure has been devised to weight the responses to the individual items in such a way as to provide the maximum accuracy of measurement.⁵ The weighted responses have been combined into a single composite *Housing Index* measuring the adequacy of each house as a dwelling place. The reliability of this *Index* has been estimated in several ways. In a selected area separate from the six communities, but typical of them, 50 houses were inventoried twice by a different interviewer, and with a time interval ranging from one to five weeks in order to estimate the errors due to the interviewer and the time of interviewing. The consistency of the scores on this test-retest survey was exceptionally high ($r = .96$), and no systematic difference was found between the two interviewers studied. Estimates of reliability by the split-halves technique of determining reliability and by the method of rational equivalence both yielded reliability coefficients in excess of .97. The accuracy of measurement reflected by these values can be seen from the following considerations: 32 per cent of the houses received their true scores, and no house received a score in error by as much as four points out of a range of 35 points.

The *Index* was validated in part by the criterion of internal consistency. The extent to which the weighting procedure automatically transformed certain first approximation

⁵ A description of the statistical technique of weighting the item-responses is being prepared for early publication.

weights which were not in accord with *a priori* considerations into weights which agreed closely with reasonable values was adduced as further important evidence of validity. A high degree of internal consistency among the *Inventory* items was revealed—houses good in one respect tended to be good in all respects, and conversely. As further validation the photographs of one hundred houses were scaled for adequacy by the psychometric method of equal-appearing intervals and the correlation between these scale values and the *Housing Index* compared. The relationship was more than satisfactorily high, ($r = .81$), but not high enough to justify substituting photographs for *Inventory* ratings. The meaning of individual *Index* scores is further made graphic by the presentation of photographs of actual houses for selected values of *Index* score (shown in the illustration Figure 1). The typical housing conditions at several score levels have been described and are presented in detail elsewhere as aids in the calibration of the *Index*.⁹

Some of the more significant descriptive findings of the housing survey are presented here. The median family consists of two adults, two children over twelve and one child under twelve years of age. Sixty-two per cent of the families own their own homes, 18 per cent rent, and 20 per cent are classified as share-croppers, squatters, or rent-free tenants. Fifty-nine per cent of the families gave farming as their only occupation; 12 per cent stated that they were on relief, or gave "public work" as their occupation.

Twenty-two per cent of the houses have only three rooms, or fewer, but 17 per cent have rooms closed off and not used, unless for storage. Thirty per cent have less than one room for each adult or equivalent; 45 per cent have no separate living room. Thirty-eight per cent used auxiliary bedrooms (rooms used during the day for some other purpose); 27 per cent sleep with two persons or more in every bed, and in at

⁹ C. I. Mosier, *Measurement of Rural Housing Status*, *loc. cit.*

least 14 per cent of the houses, sex privacy in sleeping arrangements cannot be maintained.

Forty per cent have inside walls completely unceiled, with the studding showing. Eight per cent have no decoration of the inside walls whatever, but only three per cent utilize handicraft decorations or native materials. Fifteen per cent have no pantry or storage space in the kitchen, or only poor makeshifts; 80 per cent have no kitchen sink whatever, and 52 per cent have no refrigeration whatever; 72 per cent have no electricity. Eleven per cent of the families must carry their water more than a hundred feet, and another 61 per cent have only outside hand pumps or wells. Fifteen per cent have yards littered with garbage and refuse; 16 per cent have no toilet facilities whatever, and another 70 per cent have no better than an open surface privy. The prevalence of hookworm, typhoid, and dysentery is not surprising.

Sixty-three per cent of the houses have chimneys which were judged to constitute some degree of fire hazard. Twenty-seven per cent of the houses have unglazed windows (wooden shutters only), and another 16 per cent have more than three broken panes. Only 41 per cent have all outside openings screened and in good repair to serve as protection against malaria or typhoid. The roof needs some repairing in 47 per cent of the houses, and in 6 per cent one can see daylight through it; 24 per cent show visible evidence of termite damage—only 2 per cent are termite-proofed—and 52 per cent show some degree of damage from dry-rot; 13 per cent have their foundations sagging, rotted, and crumbling. Sixty-three per cent showed no evidence of having been painted at any time; 37 per cent have no shrubs around the house and 28 per cent have no flowers; 58 per cent are lacking even bordered and sand-surfaced walks, while less than one per cent used pine-straw to surface the walks and drives.

In spite of these objective conditions, 44 per cent of the occupants mentioned no more than one aspect of the house needing repair; only 12 per cent actually planned repairs, and

MEASUREMENT IN RURAL HOUSING

the average family says that, if they won a hundred-dollar prize or found that sum, they would spend \$57.00 on the house.

The results of the survey have been tabulated for each community, and for the experimental and control groups, both in terms of the percentage response frequency for each item-response, and of the frequency distributions of the *Housing Index*.⁷ The differences between the experimental and control groups were systematically examined. Differences in housing status between the individual communities are very great—the best house in the poorest community is not as good as the average house in the best community. Differences between the experimental and control groups are small, the control group showing a slight superiority.

Photographic records have been obtained for 517 of the houses studied. These photographic records were obtained under standardized conditions, so that they may be repeated at a later date. Map records showing the location of each house in each community have been prepared and are on file. The possibility of analyzing these data to show relations with geographic factors is considered.

Achievement tests in reading, arithmetic, language, science, and mental maturity were given to 1028 pupils in all grades above the third. The results of this testing program have been analyzed by community and for the experimental and control groups. There was no discernible difference in the relative school achievement for the experimental and control groups, although there was considerable variation among the schools themselves. All schools were, on the average, markedly retarded in achievement as compared with the chronological age or the grade placement of the students. This mean retardation was from one and one-half to nearly three years, most marked in the higher grades, of course, and greater in science achievement than in any of the other fields. When,

⁷ These data are presented in full in *Measurement in the Field of Rural Housing*, loc. cit.

however, achievement is compared, not with chronological age or grade-placement, but with mental age, this apparent retardation disappeared, so that it can be said that the schools are educating the pupils to the limits of their mental capacities—assuming that the intelligence test does measure mental capacity.

The relationship between school achievement and housing conditions has been investigated. In spite of the reasons to expect that achievement would be related to the conditions of the home, the results do not bear out this expectation. The *Housing Index* showed correlations which were positive, but very low—ranging from .12 to .31—with the various measures of school achievement. The most significant relation was between *Index* and grade placement, indicating that children in the higher grades tend to come from superior homes.

Certain items of the *Inventory* were designed to measure, not the adequacy of the house, but the attitude of the family toward housing problems—insight into the housing condition, and motivation to better those conditions. When these items were studied by multiple factor analysis, the existence of a single factor of housing attitude, independent of housing adequacy, was convincingly demonstrated. This attitude variable is measured by the items dealing with willingness to spend money on the house, with ownership, with the number of repairs wanted by the occupant, with the difference between repairs needed and repairs wanted, and with whether or not repairs were planned. A method of measurement of the strength of this attitude in each family has been devised and is being applied to the individual families.

Detailed plans for a more direct attack on the problem of measuring attitudes by means of a specially designed attitude scale have been prepared. The development of this scale and its application to the families studied is a project which, it is hoped, will be undertaken at the earliest possible opportunity.

One of the contributions of this study, apart from the development of a *Housing Inventory* and its standardization,

has been the accumulation of data for subsequent analysis in connection with specific problems which will serve as the base line from which the effects of the experimental curriculum can be judged. These data—coded answers to 93 items in the *Housing Inventory* for each of 715 houses, a measure of the adequacy of each house, a record of its location and photographs for 517 of the houses, measures of school achievement for each of the 1028 children in grades 4-12 of the six schools, and summaries of the frequency and percentage frequency of each of the 685 item-responses of the *Housing Inventory* for each of the six communities and for the experimental, control and total groups—have been recorded and filed on International Business Machines punch cards, and detailed indices to this information are presented in the detailed report.⁸ How valuable it is will depend on the extent to which these data are used to provide a knowledge of the factors affecting rural housing, whether those factors be educational, sociological, psychological, or geographical.⁹

The principal results of the initial measurement program can be summarized as follows:

1. A *Housing Inventory* has been prepared and applied to the 715 homes of children in six rural Florida schools.

2. A technique for weighting these responses to the *Inventory* to yield a measure of housing adequacy, the *Housing Index*, has been developed.

3. A *Housing Index*, using the weights obtained, has been carefully standardized, using a group in addition to that on which the weights were developed. The reliability coefficient by test-retest with different interviewers was .96 and by internal consistency was .97. The *Index* has been validated by expert opinion, by internal consistency, and by comparison with psychophysical scaled judgments of adequacy based on photographs. The correlation coefficient between *Index* and scale values from judgments of photographs was .81. The

⁸ *Loc. cit.*

⁹ These data will be made available to any interested workers engaged in problems toward the solution of which these data might contribute.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

meanings of the various *Index* scores in the complete report have been interpreted by describing the typical houses and presenting photographs at various score levels.

4. Standardized intelligence and achievement tests in six fields were administered to all the children of the schools. The correlation of children's school achievement with home conditions as measured by the *Housing Index* was low for all measures of achievement, with coefficients ranging from .12 to .31.

5. By utilizing the answers of occupants to selected items in the *Housing Inventory*, a factor of housing attitude has been isolated, and initial attempts to measure it have been made.

PROCEDURES FOR HANDLING TESTS AND EXAMINATIONS

JOHN V McQUITTY
University of Florida

THE BOARD OF UNIVERSITY Examiners of the University of Florida conducts a program of testing somewhat different from that ordinarily performed by examining boards. Throughout the school year it offers a regularly scheduled series of progress tests in the basic courses in the General College. These tests are given in addition to the comprehensive examinations which are given at the completion of each course. The Board integrates the International Test Scoring Machine and punch card tabulating equipment to enable it to report results of tests promptly and adequately to all persons concerned. Also, the Board uses punch cards in building its library of test items. It is the purpose of this paper both to discuss the general work of the Board and to give the operations in detail, with special emphasis on the use of punch cards.

At the University of Florida all of the freshmen and sophomores enroll in the General College for their first two years' work. The Board of University Examiners was created in 1935 along with the establishment of the General College. The Board was charged with handling the admissions to the University and with the examinations given in the comprehensive courses which were to be offered in that college. At present the enrollment therein is about two thousand. The college examining activities of the Board come under two heads: comprehensive examinations given at the completion of the courses, and progress tests given at regular intervals of from two weeks to a month in each of these courses. The comprehensive examinations are six hours in length for two-

semester courses, and three for semester courses. The results of these examinations form the sole basis for the assignment of the student's final grade. The progress tests are similar to the comprehensives except that they cover smaller areas of the course and are usually only one hour in length. These tests are given to indicate to the student, his instructor, his parents, and the University officials how each student is doing. Even though some of these tests are given as early as eight months prior to the comprehensive examinations, the coefficients of correlation between results on progress tests and comprehensive examinations range from .65 to .83. Thus the importance of the progress tests as indicators of probable success on the comprehensive is demonstrated.

When the progress testing program was first instituted, both students and faculty were somewhat skeptical of the value of progress tests since their results were not counted when the final grades were assigned. For one thing, the practice of *not* averaging in test results at the end of the course was a new and radically different procedure and hence subject to view with considerable alarm. Now that several years have shown that the progress tests are just as important whether or not they are included in the final grade, the question of their value is no longer raised, but their usefulness is taken as a matter of course. There are two definite reasons for basing the grade entirely on the final comprehensive: 1. The grades are then assigned on the basis of how well the examinee knows the course as a whole, since piecemeal learning of the material is not enough to insure success on the examination. 2. Under such a practice the progress test results become sign posts along the way which indicate how the student and instructor are working together so that the former may achieve success on the comprehensive, but the student who compensates for an inadequate preparation and a consequent poor showing on the early tests by ultimate mastery is not penalized for his early failure.

Examinees are permitted to keep the progress test booklets, and these constitute an excellent source of material for

review. Also, the examinees' answer sheets are returned to them. Thus the student not only has a record of his raw score and his percentile rank, but also a record of the answers which he gave to the questions. By studying these answers in relation to the key of correct answers which is returned with the answer sheet, the student can make a detailed study showing which items were missed and which were answered correctly.

In addition to the tests and examinations already discussed, the University sponsors each spring a state-wide twelfth-grade high-school testing program, the results of which are used by the University as entering placement tests. The General College handles the announcements of the program and the distribution and receipt of the test materials.

Separation of Instructing and Examining

In theory there is complete separation of the teaching and examining functions. The examining and the issuing of final grades are both done by the Board of Examiners. However, in actual practice there is the closest cooperation between the instructional staffs and the Examiners. In most of the courses the construction of the test items is done by persons engaged in teaching those courses. These items are then subjected to critical review by the Examiners, and any test items composed by the Examiners are reviewed by members of the instructional staff. In all cases the test or examination has the approval of both the instructional staff and the Examiners. All tests are printed, given, and scored by the Examiners. When it comes time to set the grades—i.e., determine the raw-score division points for the passing grades A, B, C, D, and the failing grade E—the members of the instructional staff cooperate again. The members of the staff and a representative of the Examiners hold a meeting scheduled for this purpose. In this task, use is made of all pertinent objective information about those students making a given raw score; the results of the entrance examinations and of the progress tests throughout the year as well as the characteristic responses to those

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

test items considered most crucial are all considered for students at the critical values dividing letter-grade equivalents. Also at hand are the distribution of raw scores on the examination and the corresponding percentile ranks. The anonymity of the individual student is preserved throughout this procedure. Again it must be made clear that all these data are used as aids in determining where the grades should come on the distribution of raw scores. In no sense are any of the data "averaged-in" when the grade is assigned. Once it is decided, for example, that scores of 400 or above are to receive A's, everyone in that category—but no one else—receives an A, and so on for the other grades. There is no "grading on the curve" in the sense that a predetermined distribution of grades is followed. This is shown by the fact that even in a course taken by as many as 700 persons the percentages of A's has varied from 5 to 11 and the percentage of failures from 12 to 21. Since 1935 the Board of Examiners has assigned 42,214 final grades with the following distribution:

TABLE 1

DISTRIBUTION OF GRADES FOR COMPREHENSIVE EXAMINATIONS
WINTER, 1936 THROUGH SUMMER, 1941

Per Cent for Each Grade*					Total	Per Cent	Total
A	B	C	D	E	Examined	Absent	
8.42	16.72	37.35	21.85	15.66	41,112	2.68	42,214

*Based on number examined.

Office Routine

In discussing the routine for handling the test results and the test items, special emphasis will be given to those procedures which may not be widely known. It is recognized that practices for the processing of examinations will vary according to the use to be made of the test materials and with the mechanical equipment available. The Board of Examiners makes use of the following mechanical equipment supplied by the International Business Machines Corporation: test scoring machine with graphic item counter, alphabetical printing punch,

TESTS AND EXAMINATIONS PROCEDURES

high speed reproducer, interpreter, collator, and alphanumeric tabulator with 25 alphabetical and 30 numeric type bars. The routine of handling the progress tests is affected by the facts that all answer sheets are to be returned to the students, and that the sheets carry the raw scores and the percentile ranks. In the case of the comprehensive examinations the answer sheets are retained by the Examiners, the examinee receiving nothing but his letter grade. Also, the need for prompt reporting of results is particularly great in the case of progress tests because it is felt that the results are more helpful to the students if received while interest is still high. Hence, progress tests are usually given on a Saturday morning and the results returned to the instructors, administrative offices, and the students the following Monday, even though as many as 1000 students are given two tests each. In the case of the comprehensive examinations there is an equally great need for prompt work because all examining for the year's work for the freshmen and sophomores must be accomplished and final grades submitted within a period of two weeks. It would be impossible to do all this on the limited budget of the Examiners without making full use of the punch-card system. The integration of punch-card methods with the test scoring machine, however, permits a very small staff to handle a large volume of tests in a short time and to make the results available in a variety of forms to meet the varying needs of student, instructor, department head, and administrator. The use of punch cards will be discussed under the four following heads: (1) placement tests; (2) progress tests; (3) comprehensive examinations; and (4) library of used test items.

A Punch Cards and Placement Test Results.—In the spring of 1941 the following tests were used in the high-school testing program:

1. The Henmon-Nelson Test of Mental Ability, Form B
2. Cooperative English Test, Effectiveness of Expression, Lower Level, Form Q

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

3-5 Cooperative Achievement Tests, Form QR, in

3. Social Studies
4. Natural Sciences
5. Mathematics
6. Cooperative French Test
7. Cooperative Latin Test
8. Cooperative Spanish Test

All of these tests are given with separate answer sheets which are machine-scored by the Board of Examiners. All of the test results are punched into tabulating cards, and the following steps are employed in the procedure:

- 1a. Answer sheets are handled by schools, and when the sheets are received, they are separated by tests and alphabetized. Some visual check such as a corner-cut or a punched hole will aid in the checking of the homogeneity of a pile of answer sheets.
- 2a. Name cards are punched for each examinee. Each card carries a code number indicating the high school. Sex is indicated by using F for female and M for male. A heading card with a characteristic control punch is made for each high school.
- 3a. The name cards are listed on the tabulator in alphabetic order by high schools on a prepared form which carries eight columns, one for the raw scores for each test.
- 4a. The answer sheets which have already been separated according to tests and alphabetized are checked against the list described in 3a, to make certain that the answer sheets are in the same order as the names on the list. In the case of persons who took one but not all of the tests, the areas where results for the missing tests would be recorded are marked. The practice of having the answer sheets and names in identical order, with absentees designated, facilitates the recording of the machine scores.
- 5a. The answer sheets are scored on the test-scoring machine and the raw scores recorded in the appropriate rectangle on the name lists. Usually two persons are used, a machine operator and a recorder. The operator gives the scores orally to the recorder, who enters them in the proper place.
- 6a. The answer sheets are scored again and the raw scores checked against those recorded in number 5a.
- 7a. The checked raw scores are punched into the name cards. The name cards and the lists carrying the raw scores are in the same order.

TESTS AND EXAMINATIONS PROCEDURES

- 8a. The name cards are listed on the alphanumeric tabulator to show name, sex, and raw scores. A comparing control is maintained on high-school code number.
- 9a. The lists in number 8a are checked orally against the lists on which the raw scores are written.
- Note:* From now on all operations are entirely mechanical.
- 10a. As soon as all of the scoring and punching has been done, the distributions of raw scores are made by sorting the cards in order by raw scores and running them through the tabulator with a comparing control on raw scores if an interval of one is desired (if any other interval is desired, it is necessary to use interval heading cards to establish the control breaks); and progressive totaling is used to secure the cumulative frequencies.
- 11a. Percentile ranks are computed for the distributions obtained in number 10a. The percentile ranks are punched into heading cards which carry raw scores and corresponding percentile ranks. The corner-cut on the heading cards should differ from that on the score cards.
- 12a. These percentile rank cards, which must carry an appropriate control, can then be collated into the raw-score detail cards. (Steps 10a, 11a, 12a, and 13a should all be done for one test before anything is done for another test, if the sorting is to be kept to a minimum.)
- 13a. By using the high speed reproducer the proper percentile ranks are punched from the heading cards into the detail cards, using a control punch to clear the punch magnets at the proper time.
- 14a. After all of the gang punching has been done (all gang punching should be sight checked and the detail cards checked on the collator for proper sequence order), all percentile ranks can be interpreted at one time.
- 15a. The detail cards, which now carry percentile ranks, are sorted alphabetically on three letters, then sorted by high schools using the high-school code. Then the detail cards are checked by hand to insure correct alphabetical order.
- 16a. Lists are run for each high school, showing all percentile ranks for each examinee. Under some conditions it may be desirable to run master alphabetic lists before the detail cards are sorted by high schools.
- 17a. A master list of the results for all examinees in alphabetic order is prepared on the alphanumeric tabulator. Usually this list is prepared on a stencil or duplicator paper, so that a large number of copies can be made.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

B. *Punch Cards and Progress Test Results.*—The nature and use of progress tests has already been discussed. Punch cards are used here as an aid in making the results available quickly to the instructors and administrative officers. The steps in preparing the cards and using them are as follows:

- 1b. As soon as registration is complete, the class cards in the Registrar's Office are duplicated for each course in which progress tests are to be given. The information picked up from the Registrar's card is:
 - (1) Student number
 - (2) Student name
 - (3) Course and section
- 2b. The cards prepared in number 1b are alphabetized for three letters on the sorter and the alphabetizing checked by hand.
- 3b. The collator is now used to insert a sequence number card in front of each set of cards for the same student. These sequence cards are prepared in advance and carry a control punch as well as numbers in sequential order from 0001 to 3999. Only odd numbers are used, the evens being reserved for future expansion due to errors or to late registration changes.
- 4b. The progress test cards with the inserted sequence cards are run through the tabulator with a comparing control on both sequence number and student number, with the machine set to tabulate. The resulting list is then checked visually to see that a sequence card has been inserted at the proper place and that the test cards are in proper alphabetical order. Any errors are corrected.
- 5b. The deck of cards used in number 4b is run through the reproducer, and the sequence numbers are punched from the sequence cards into the test cards.
- 6b. The progress test cards can always be placed in strict alphabetical order merely by sorting them on the numerical sequence number of four digits.
- 7b. Next, placement test deciles are gang punched into the test cards. This is done by sorting on student numbers (both the placement test decile cards and the tests cards carry student numbers), with decile card coming first, and running the entire deck through the reproducer with a control on a suitable punch in the decile card.
- 8b. When a progress test is given, the answer sheets are scored on the test scoring machine, the answer sheets distributed, and percentile ranks computed and recorded on the answer sheets. Letter grades are recorded also, if any are assigned.

TESTS AND EXAMINATIONS PROCEDURES

- 9b. The answer sheets are alphabetized and checked against the deck of progress test cards for that course, cards are pulled for absentees; and both the answer sheets and the cards are put in the same order to facilitate punching the test results into the cards.
- 10b. Percentile ranks (and grades, if any) are punched from the answer sheets into the test cards on the alphabetical printing punch. The punching is checked.
- 11b. The cards now go to the tabulating department where the following operations are executed:
 - (1) Absentees are re-inserted on the collator and the sequence of the cards checked on the collator.
 - (2) An alphabetical list of all students is prepared on the tabulator showing the following data: student name and number, course and section, and the results of all progress tests to date.
 - (3) An alphabetical list of students by sections is run on the tabulator, showing the same data as for (2), above.
- 12b. The lists made in (2) of 11b are checked orally against the answer sheets as an added precaution to insure accuracy.
- 13b. The answer sheets and lists of results are given to the proper instructor for each section. The lists are for his use; the answer sheets are returned to the students. Also, the students receive key sheets of the correct answers, so that they may see just what they missed on the test.
- 14b. About once a month a composite list is run on the tabulator. This list shows the progress test results for all courses for each student, and through them it is possible to see how the student is doing in all of his courses. The sequence number is used to put the cards in one alphabetical order, where all the cards for each student are together. It has been found best to keep the cards in the order in which the composite lists will be run. When a test is given, the cards for that test only are selected from the entire deck. As soon as the details of handling that test are completed, the cards are collated back into the composite deck.
- 15b. By the end of the school year, the cards represent a complete picture of the record of each student in each course. All kinds of statistical studies are possible from these cards, among which are:
 - (1) Correlation between placement tests, progress tests, and comprehensive examination grades.
 - (2) Investigation of quality of work done by those who drop or resign.
 - (3) Correlation between grades in different courses.
 - (4) Grade distributions

C. *Punch Cards and Comprehensive Examinations.*—At the end of the school year it is necessary to give, score, and report grades for all the comprehensive examinations within a period of less than two weeks. It has been found that the use of punch cards speeds the work and increases the accuracy. The steps in preparing and using these cards are:

- 1c. A deck of comprehensive examination cards is made for each course by reproducing the progress test cards, except that it is necessary to omit the first two progress test results to make room for the raw scores on the comprehensive. The remainder of the card is reproduced because the placement test results and the progress test results are useful as aids in setting the comprehensive grades.
- 2c. Decks of master cards for raw score intervals are prepared. These are used to enable the tabulator to make the distribution of raw scores for each examination. Decks with intervals of 2, 5, and 10 have been made. It has been found helpful to have duplicate decks to facilitate handling of courses where identical intervals are used. These cards carry a control punch.
- 3c. Percentile rank master cards are prepared also. An interval of 1 is used, and the range is from 01 to 99. It is well to have several sets of these and to have an abundance of cards for ranks below ten and above 90, since several class intervals may have the same percentile rank within these ranges. These cards carry a control punch and are used to gang punch percentile ranks into the examination cards.
Note In all instances it is well to have the master cards with a corner-cut different from that of the detail cards.
- 4c. In handling comprehensive examinations the student number rather than the name is used to identify the student. This is done to impersonalize the examining and to simplify the procedures, because operations can be done more readily on a numerical than on an alphabetical basis.
- 5c. These cards are used to prepare attendance lists for each examination room and a master-list for use in checking in the papers at the end of the examination.
- 6c. As soon as the examination is over, the answer sheets are placed in numerical order and checked against the check-in rolls. This is done to make certain that no answer sheets have been misplaced. A special form is filled out for each absentee and inserted in its proper place in the stack of answer sheets. This means that there is either an answer sheet or an absentee sheet

TESTS AND EXAMINATIONS PROCEDURES

for each name on the roll and for each examination card. This has been found to be preferable to pulling the cards for the absentee.

- 7c. The answer sheets are scored on the test-scoring machine. If more than the front of one answer sheet is used, all of the raw scores are recorded on the front of the first answer sheet. If the student has more than one answer sheet (which is usually the case, since most of the examinations have both morning and afternoon sessions), great care must be used to be certain that all the scores on the sheet are for the same individual.
- 8c. After the answer sheets have been scored and checked and the addition of the scores completed and checked, the total raw scores are punched into the cards mentioned in number 1c. The punching is facilitated because both the cards and answer sheets are in order by student number and there is either an answer sheet or an absentee blank for each card.
- 9c. From now on all operations are mechanical except the checking of the punching of the total scores. After the checking, the cards are sorted on total score, with the interval cards being placed first in the hopper. This sorting places the cards in order by score, with the interval cards coming at the proper place. It is well to check the sequence of the cards on the collator.
- 10c. The cards are now tabulated with a control being taken on the 11-zone punch in the interval cards (the detail cards being blank in the control column). The tabulator will print the intervals from the interval cards, count the detail cards between interval cards, and record the frequencies and the progressive totals. To insure that the count for each interval is placed on the same line with the intervals, only the upper hub of the comparing control must be used (i.e., there is no plugging to the add-hub of the control source), so that a control change will occur only when going from blank to X-punch, but there will be no change when going from X-punch to blank.
- 11c. The percentile ranks are computed from the progressive totals, and the proper percentile rank card (selected from the prepared deck of percentile rank cards) is inserted manually just back of the interval card with which that percentile rank goes.
- 12c. The cards are re-tabulated with both the interval cards and the percentile rank cards in the deck. In this operation the percentile rank card is handled as a detail card (i.e., it is blank in the control column), but it carries another X-punch to keep the tabulator from counting it. By taking the percentile rank through a counter, it is possible to print the ranks on the same line with the interval, the frequency, and the progressive total. Both interval cards and detail cards are blank in the percentile

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

rank column, so that all the counter receives is the percentile rank. This re-run serves as an accurate check on the original distribution and the insertion of the percentile ranks at the proper place. The first step of the check is to compare the second distribution with the first to check frequencies and progressive totals. The insertion of the percentile rank cards can be checked by comparing the machine-recorded ranks with those obtained in the computation on the first distribution.

- 13c. The complete deck of cards is run through the reproducer to gang punch the percentile ranks into the detail cards. It is not necessary to remove the interval cards before this operation if both the interval cards and the percentile rank cards contain a common X-punch which can be used to clear the punch magnets. It should be recalled that the percentile rank cards were inserted behind the interval cards. After the punching has been finished and sight-checked, both the interval cards and the percentile rank cards can be separated from the deck by using the sorter. Then the percentile ranks are interpreted.
- 14c. The deck of comprehensive examination cards is left in order by score and percentile rank until after the grades are set. Since the cards contain the placement test results, most of the progress test results, and the score and rank on the comprehensive examination, the information they reveal helps in setting the grades. Also, the distribution of raw scores and the specified answers given to certain key questions are used in setting the grades. For example, in setting the grades someone may wonder what kind of persons we find at the 10th percentile from the bottom. By referring to the comprehensive examination cards, we can learn the quality of their placement test results and their relative success on progress tests, and by referring to the answer sheets we can see which questions they missed and which they answered correctly. If we find that most of the students at this percentile are missing items based upon elementary facts and principles, we feel that we cannot pass persons at that level. A higher level can be investigated until a satisfactory one is found. Such a procedure can be used until all of the division points for the various grades have been set.
- 15c. After the grades have been set, the grades are gang punched into the detail cards and the grades interpreted.
- 16c. Next the detail cards are alphabetized on the sorter by sorting on sequence number, and grade reports are printed by the tabulator on specially prepared forms, so that copies of the grades can be reported by the Registrar and to others concerned.

TESTS AND EXAMINATIONS PROCEDURES

- 17c. Since the grade cards contain so much pertinent information regarding each student's academic record during the year, many statistical studies can be made from them.

D. *Punch Cards and the Library of Test Items.*—An item analysis is made of the items used on each progress test and comprehensive examination. In making the analysis, 100 answer sheets uniformly distributed throughout the highest and lowest quarters are selected as a sample. (If the number of examinees is less than 400, samples of 50 may be used, or the upper and lower halves may be used instead of quarters.) The analysis is made on the test-scoring machine by utilizing the graphic item counter, and the count is made on the correct responses. For the items found to have low discriminating value the frequency counts for each distractor are made, unless the low validity is obviously due to excessive ease or difficulty. Four measures are secured for each item: V, the validity or discriminating power, which is the tetrachoric coefficient of correlation between the item and the total test; D, the difficulty, expressed as per cent of the group answering the item correctly; H, the per cent of the highest quarter or upper half answering the item correctly; and L, the per cent of the lowest quarter or lower half answering the item correctly.

An outline is made of the course content of each comprehensive course with major, intermediate, and minor classification of topics, and on the basis of this outline each test item is classified according to the aspect of the course which it covers. This classification and the item analysis data are all punched into cards, and the statement of each item is typed on the back of the card. This provides a very complete and usable library of test items. The actual steps in preparing the cards for this library are:

- 1d. Pre-punching a deck of cards to show the course and the date the test was given.
- 2d. Punching the validity data (V, D, H, and L) and the course content into the cards prepared in 1d. One digit is used for each validity datum, i.e., a validity correlation coefficient of .30-.39 is written as 3 and for the per cents for D, H, and L

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

the units value is dropped; for example, a per cent of 60-69 is punched as 6. This is done to conserve columns on the cards. (Of course, this procedure will mean that the punched values will average about .05 or 5 lower than the computed values.)

- 3d. The test item involved is typed on the back of the card which carries the corresponding item analysis data.
- 4d. The typing is proofed, the correct answer indicated, and the item is filed according to test content.

The information which is punched into these cards makes it possible to select mechanically items according to validity, difficulty, and course content and to make various types of statistical studies.

MACHINES IN CIVIL SERVICE TESTING

SIDNEY W. KORAN

Employment Board, Pennsylvania Department of Public Assistance

EDITOR'S NOTE: This article is an abridged version of a lecture recently presented by the author before an in-service training class comprising staff members of the examination division of a large state civil service agency. It is offered here because it pulls together, for what is probably the first time, numerous loose ends of the important body of knowledge that is beginning to come into existence on the mechanization of civil service examining processes. The article comprises a description of the purpose, design and operation of the I.B.M. scoring machine, a discussion of the limitations of the scoring machine in connection with the conduct of examinations, information on adapting tests to machine scoring, descriptions of procedures for scoring tests using the I.B.M. and other machines, information on scoring various types of rating scales by machine, material on the uses of the scoring machine in item analysis and in the computation of several statistical measures, and a summary of the place of tabulating equipment in the conduct of certain examination tasks.

The I.B.M. Scoring Machine

The I.B.M. scoring machine was designed to meet a very real and pressing need in the field of educational and psychological testing for a method of scoring objective tests that would combine speed, accuracy, and low cost to an extent considerably beyond that possible with any technique previously developed. The recognition of that need was sufficiently great to stimulate, over a period of years, the development of numerous ingenious techniques as well as several different types of automatic and semi-automatic devices. The one, however, which appears to have been the most successful and to have earned the widest application to civil service use is the comparatively new I.B.M. scoring machine.

The operation of this machine is dependent upon the principle that the mark made by a pencil having a soft lead will conduct electricity. In order to score a test by the machine it must be designed so that the examinee may indicate his answers to the questions by placing pencil marks in certain predetermined and properly labeled positions on a sheet of paper which is either separate from the test booklet or may later be separated from it. For consistently satisfactory results it is advisable to furnish examinees with mechanical pencils equipped with special high-graphite-content leads and to use answer sheets that have been carefully and accurately printed so that the location of each one of the 750 possible response positions on either side of the sheet corresponds within fairly close limits to the location of each of the 750 sets of contacts within the machine.

Each of these sets of contacts consists of five small parallel blades insulated from one another and connected alternately to the positive and negative sides of the electrical circuit, the current for which is furnished by several conventional radio "B" batteries. When an answer sheet is inserted in the machine for scoring, it is pressed against a plate containing the 750 sets of contacts. Whenever one of these sets of contacts presses against a pencil mark, the latter, since it is a conductor, permits current to flow across one or more of the four gaps made by the five parallel blades. The length of the pencil mark determines whether one, two, three, or four of these gaps will be bridged. If the examinee follows instructions and makes his pencil mark sufficiently long, it will bridge all four of the gaps. By designing the machine so that there are several millions of ohms of resistance in series with each set of contacts, current differences which result when some of the examinee's marks are not long enough to bridge all four gaps are minimized sufficiently to prevent their having an appreciable effect on the score.

The resistance in each circuit is such that when the appropriate rheostats have been adjusted properly, a single unit of

CIVIL SERVICE TESTING

current is registered for each set of contacts that may be pressed against a pencil mark. Scores are read on a meter which has been calibrated in terms of these units. The use of switches and other accessories makes it possible to read rights, wrongs, omits, rights minus wrongs, rights plus wrongs, rights minus or plus a fraction of wrongs, etc. Whether any given choice, or answer position, will be counted as right or as wrong, or whether it will be eliminated from the scoring altogether, is determined by the manner in which holes have been punched in the set of keys inserted in the scoring rack. By the use of switches and the proper perforation of field selection holes in the scoring key, the machine may be adjusted so that the meter will read the score for all of the items on the answer sheet or for certain combinations of the ten 15-item fields, or both.

Machine-Scorable Answer Sheets

Standard answer sheets designed to fit several types of general situations are available from the manufacturer of the machine, and it is also possible to have special answer sheets, designed to meet certain specific requirements, printed to order. Unless ordered in fairly large quantities, however, special answer sheets are usually more expensive than standard answer sheets and their use frequently introduces an additional time element into the planning of examinations.

Some of the agencies, in an effort to take advantage of the economies afforded by quantity purchases and to enjoy the convenience of having their own stock on hand, have standardized their major requirements sufficiently to permit them to order relatively large quantities of three or four types of answer sheets printed only with the name of the agency, the item numbers, and the response positions. As new examinations come up, these agencies select the type of answer sheet which most nearly fits the requirements of the particular situation and print or multilith, in the left-hand margin, whatever additional identifying material is required.

Limitations Imposed by Machine Scoring

To the individual who is about to construct a test that is to be scored by machine the limitations imposed by the machine method are chiefly three: 1. A separate answer sheet must ordinarily be used. 2. The response to each question must be indicated by making a special kind of pencil mark. 3. The orientation of the response positions on the answer sheet cannot be altered.

The first of these considerations, that of the use of a separate answer sheet, is more correctly a "condition" rather than a "limitation," for even when the machine method of scoring is not involved it is usually desirable, when examinations comprising large numbers of items are to be administered to any considerable number of individuals, to make use of some form of separate answer sheet in order to facilitate the scoring process. This is true whether it is planned to do the scoring entirely by hand or by some combination of hand-scoring and overprinting (with a multilith, for example).

Another reason why the use of a special answer sheet is not ordinarily a serious obstacle is that it is frequently possible, if necessary, to design the examination so that the test questions are printed directly on the answer sheet beside (or directly over or under) the response positions. This has been done in the case of a number of standardized educational and psychological tests and has also found some, though much more limited, use in connection with civil service examinations. While this procedure seems to be particularly advantageous when used with one- or two-page personality inventories and, as will later be pointed out, certain types of rating scales, there are ordinarily several objections to its routine use in setting up civil service examinations. Among the principal objections are the increased trouble and expense caused both by the special printing requirements and by the fact that the relatively small number of items which may be printed on a letter-size sheet usually necessitates using several answer sheets for a test of any appreciable length. The handling, scoring, and computational difficulties encountered whenever a single test requires more

than both sides of one answer sheet are ordinarily sufficient to discourage that practice. There are, however, certain situations in which the mere fact that several answer sheets will be required for a given test may well be a matter of relatively minor importance when considered beside the larger aims of the examination.

The second limitation imposed by the use of the scoring machine method, that of the necessity for making a special kind of pencil mark to indicate the answer to each question, is apparently proving less of a problem than many expected it would. As returns come in on the results of research (1) and on the development of improved scoring procedures which have reduced the likelihood of scoring errors to a negligible figure (3, 11), it is clear that whatever problem is actually presented by this limitation may be pretty adequately neutralized by taking the following three steps:

1. Include in the examination announcement a section comprising (a) an explanation of the types of questions that will be used, (b) a statement of the fact that the test will be scored by a machine which will provide the correct score if the examinee follows all instructions carefully, (c) a list of the rules which must be followed in indicating the answers, and (d) a group of sample questions printed alongside a specimen portion of the answer sheet on which the answers to some of the sample questions have been properly marked and the remainder left for the examinee to complete as a practice exercise.¹

2. Provide the examinee, at the time the examination is administered, with a page of instructions, similar to those described above, which he is given time to review sufficiently long to refresh his memory and to which he may refer at any time during the examination. (One example of such an approach is the *Directions for Using the Answer Sheet*, reproduced on the following pages, which has been used in Pennsylvania since 1939 by the Employment Board of the Department of Public Assistance with examinations conducted for approximately 115,000 persons.)

3. Furnish, throughout the test, additional adequate and clear directions including, wherever a somewhat different approach is employed, a sample question properly answered (7). (Several illustrations of such special directions appear among the examples of test material presented later in this article.)

¹ An example of this approach is the page entitled *Sample Questions for the General Test* which appears as part of current U. S. Civil Service Commission announcements for examinations which will include machine-scored written tests.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

DIRECTIONS FOR USING THE ANSWER SHEET

All of the answers in the test you are about to take are to be recorded on special ANSWER SHEETS instead of in the Question Booklet. *To receive credit for your answers they must be recorded in the proper spaces on your ANSWER SHEET.*

ANSWER SHEETS will be scored by an electrical test-scoring machine. In order for your test to be scored accurately, it is necessary for you to observe the following directions carefully:

1. Read each question and its numbered answers and decide which answer is correct.
2. Find the pair of dotted lines numbered the same as the answer you have chosen as being correct, and blacken this space with your pencil. *Be sure that the space you mark is in the row numbered the same as the question you are answering.* Misplaced answers are counted as wrong answers.
3. Indicate each of your answers with a vertical solid black pencil mark. Solid black marks are made by going over each mark two or three times and by pressing firmly on your pencil.
4. If you change your mind, *erase your first mark completely, then mark the correct space.* Blacken one space *only* for each question number.
5. Do not rest the point of your pencil on the ANSWER SHEET while you are considering your answer and *do not make unnecessary marks.*
6. Keep your ANSWER SHEET on a hard surface while marking your answers.
7. Make your marks as long as the pair of dotted lines.

Below are some sample questions to give you practice in using the ANSWER SHEET. The questions at the left are similar to those you will find in your Question Booklet. At the right is an illustration of a portion of an ANSWER SHEET. The answers to the first four questions have already been marked on the ANSWER SHEET. Study the questions and note the way the answers to them have been marked on the ANSWER SHEET. Then answer each of the remaining questions in exactly the same way; that is, by making a heavy black mark on the ANSWER SHEET in the space numbered the same as the correct answer.

Questions for Practice

1. The third month of the year is: (1) February, (2) March; (3) January.
2. The capital of Pennsylvania is: (1) Harrisburg; (2) Albany; (3) Boston.
3. The Governor of Pennsylvania is: (1) John Garner; (2) Alfred Landon; (3) Arthur James.
4. If one pencil costs one cent, five pencils will cost: (1) three cents; (2) six cents; (3) four cents; (4) two cents; (5) five cents.
5. George Washington was the first President of the United States. (1) True, (2) False.

	1	2	3	4	5
1					
2					
3					
4					
5					
6					
7					
8					

CIVIL SERVICE TESTING

6. Every calendar year has: (1) eleven months; (2) ten months; (3) twelve months.
7. The fuel most commonly used in automobiles is: (1) kerosene; (2) carbona; (3) crude oil; (4) gasoline.
8. The sum of six and four is: (1) five; (2) six; (3) eight; (4) nine; (5) ten.

NOTE. The Answer Sheet provides spaces for recording five different choices for each question. Some of the questions in your examination booklet may contain only two, or three, or four choices. When answering a question containing fewer than five choices, you are to ignore the additional spaces printed on the Answer Sheet for that question.

The last of the three limitations mentioned as imposed by the machine method of scoring has to do with the fact that the relative location of each of the 750 response positions on the answer sheet is fixed and may not be changed by the test constructor. It is in meeting this difficulty that the test technician's ingenuity enters the picture.

Despite the handicaps of perpetually imminent deadlines and understaffed examination units—two well-known characteristics of the conditions under which many civil service commissions work—a sufficient number and variety of adaptations of test material to this particular limitation of the machine method have been produced in the short space of a few years to warrant the conclusion that the use of separate answer sheets involving fixed response positions offers no particularly serious obstacle to the construction of objective test material. In fact, what used to be a double bottleneck—construction and scoring—has, because of the advantages offered by the scoring machine, been reduced to but a single obstruction. Objective tests may now be scored so cheaply that the major remaining obstacle to the wider use of *better* objective tests appears to be the difficulty of constructing them under the usually prevailing conditions of insufficient time and the not-too-ready availability of trained examination technicians.

Constructing Items for Machine Scoring

In general, the basic rules to be followed and the pitfalls to be avoided in the construction of good objective-type items

are as applicable to items that are to be scored by machine as they are to items that are to be answered either directly in the test booklet or on a separate sheet designed to be scored manually. An item that is "tricky" or that contains an ambiguous or ludicrous statement is unacceptable for reasons quite apart from the scoring method that will be employed. While there are certain additional points that must be kept in mind—mostly with regard to adequate instructions to the examinee and strict adherence both to the physical limitations of the answer sheet and the electrical limitations of the machine itself—fundamentally, an item that is unsatisfactory for any reason that would interfere with its suitability as an ordinary objective-type question is equally unsatisfactory for use in a machine-scorable test. Here, however, are some of the considerations which appear to be sufficiently peculiar to the use of machine-scorable answer sheets to warrant enumeration and brief discussion:

1. *Choice of answer sheets.* Wherever practicable, test items should be designed to make use of standard answer sheet forms. Doing so keeps down construction time as well as costs and obviates the necessity for presenting special instructions not covered by the general directions printed in the announcement and furnished the examinee at the time of the examination. The use of standard answer sheets possesses the additional advantage of capitalizing on the fact that, since machine-scored tests are being used more and more widely by civil service agencies and educational institutions, an increasingly large proportion of the civil service test-taking population may be expected to have sufficient previous experience with standard forms of separate answer sheets to permit them to concentrate on the test material with a minimum of distraction and tension.

2. *Adequate instructions.* Instructions for indicating answers to such specialized subtests as those involving alphabetizing, proof-reading, checking, sorting, filing, punctuation, and similar tasks should be adequate and should preferably include a sample exercise properly answered. In writing these

directions the kind of language, specificity, and need for examples will, of course, vary according to the level of the job for which the examination is being designed. In general, however, it is safer to be somewhat too specific and to provide what, to the sophisticated test-taker or Ph.D. test constructor, may sometimes appear to be an unnecessary example, than to take too much for granted on the part of the examinee.

It is sometimes argued that an examinee who can't follow such simple instructions "doesn't deserve to pass" or "couldn't do the job anyway." While there are certainly times when this stand appears justifiable, the writer's opinion is that it is always safer to provide, if for no other reason than the maintenance of satisfactory public relations, directions that meet the highest standards of adequacy. If it is desired to test the examinee's ability to follow instructions, one should use a test designed to do just that, rather than take the chance of measuring such a trait by using "complicated" (in this case, a euphemism for "inadequate" or "unsatisfactory") instructions which are likely to result in a distribution of scores unduly influenced either by the degree of an examinee's previous experience with new types of tests or by the extent to which he exhibits caution in situations of this kind.

In this connection it sometimes occurs, in construction of a test for a position such as building superintendent, that by the time the test constructor has finished adapting a certain bit of practical material to the limitations of the answer sheet, his product, regardless of how cleverly worked out, is no longer suitable for that particular level of job. What he has may be an excellent test of intelligence, but of a higher level than that required of a building superintendent. It hurts, sometimes, to have to discard or extensively modify a brain child of that kind, but it has to be done.

3. *Item sequence.* The sequence of items in the test should be such that subtests or item-groups that are to be weighted differently from other portions of the test or for which a separate score will be desired, will fall entirely within one or more fields. If, for example, in a test for key punch operators

a separate score will be required for a 30-item subtest on the subject of coding, the items for the entire test of which the coding items are a part should be arranged so that the item numbers (if a standard answer sheet is used) will start with 1, or 16, or 31, or 46, etc. When the test is being scored it will then be possible, if the proper field selection holes have been punched in the answer key, to read the score of the subtest with the expenditure of no more additional effort than is required for turning a knob while the answer sheet is in the machine. Similar treatment should be accorded item groups to which a correction formula is to be applied that is different from that used for any other part or parts of the entire test.

4. *Reducing response errors.* Care should be taken to avoid wording questions and selecting styles of type or print that are likely to cause the examinee to make unnecessary clerical errors in recording his responses. This not infrequently occurs, for example, when Arabic numerals having the same range as those used to denote response positions are used in the answer. Answer sheets are available which eliminate this difficulty by using the letters A, B, C, D, and E to designate the response positions. Where the response positions are numbered, however, it is frequently helpful simply to spell out the numbers from 1 to 5 when they appear alone or almost alone in the answer. Two simple illustrations of this point are:

"The number of inches in one-third of a foot is: (1) two; (2) three; (3) four; (4) five; (5) six." *instead of*.
 "(1) 2; (2) 3; (3) 4; (4) 5; (5) 6."

"How many persons in the family are eligible to receive some form of assistance? (1) none; (2) one; (3) two; (4) three; (5) four." *instead of*: "(1) 0; (2) 1; (3) 2; (4) 3; (5) 4."

5. *Juxtaposition of instructions and related items.* Where use is made of a page of instructions including a key, legend, code, or similar device likely to have to be referred to frequently by the examinee in order to answer a given group of related questions, the format of the examination booklet should be such that the page containing the instructions will

face at least a group of the questions. Among the possible exceptions to this rule is the situation in which one of the functions being tested is the ability to memorize certain material or relationships, and in which the test is being timed in an effort to obtain a measure of the examinee's ability to do so.

6. *Completion arithmetic items.* The construction of choices for items involving arithmetic, algebraic, or statistical problems, or consulting a graph or chart, is no different when the item is to be scored by machine than by any other method. There may, however, be situations in which it is considered important, in connection with a certain group of items in a test, to know the *exact* answer arrived at by the examinee as a result of his calculations. When this is so, it is possible by expending some additional time and effort, to retain the advantages of the completion type test for this particular group of questions and at the same time have the machine-indicated score represent the examinee's achievement in the entire test. This may be accomplished by designing the answer sheet so that spaces are provided both for the examinee to write in his answers to the questions and for a clerk to indicate the correctness of those answers by making the usual kind of pencil marks in response positions especially provided for that purpose.

Two disadvantages of this approach are the need for special answer sheets and the time and expense involved in having the completion items scored manually before the test as a whole can be scored by machine. Another possible disadvantage is the effect that the use of two types of directions may have on the examinee's adherence to the important and oft-repeated general instructions to "indicate your answer to each question by making a heavy mark in the appropriate space on the answer sheet." This is of some importance, for the extent to which the examinee can be persuaded to accept the idea of making proper marks instead of writing answers or numbers on the separate sheet is one measure of the amount of machine vs. handscoring that will have to be done. For these reasons, and because it is probably possible to use the multiple-

choice form of presentation for arithmetic and similar items without interfering seriously with their validity, it would seem preferable, ordinarily, to avoid mixing the two types of responses in the same test.

Beginning on the next page are examples of Test Material Adapted to Machine Scoring.²

Scoring Civil Service Tests with the I.B.M. Machine:
Historical Note

Civil service commissions, while recognizing the speed and economy features of the machine-scoring method of rating objective tests, were at first quite reluctant to put the machine to use. Springing in part, probably, from the usual resistance to adopting methods and procedures differing from those already in use, the criticism was made that not only did the machine method involve the use of special materials to which the test-taking public might object, but the scores which it turned out were insufficiently accurate.

This controversy was quietly running its course when an event occurred in the field of public personnel administration that resulted, among many other things, in removing the whole question from the *talking* to the *trying* stage. Suddenly, all over the country, sizeable civil service agencies were coming into existence in accordance with the merit system provisions of the Social Security legislation. Many of these new commissions were faced with the task of examining unprecedented numbers of persons within time and budgetary limitations that were not easy to meet, and some were composed of commissioners and administrators who were sufficiently new to civil service problems to be relatively quite receptive to such new-

² These examples of machine-scorable subtests and item-groups are offered solely for the purpose of illustrating the variety of test material that may be adapted to machine scoring and the kinds of instructions that may be employed. The writer wishes to thank the Employment Board of the Pennsylvania Department of Public Assistance and Miss Hilda P. Thompson, Executive Director, for their kind permission to use this material, which was developed for the Board over a period of several years by C. R. Adams, G. K. Bennett, S. W. Koran, B. V. Moore, E. A. Rundquist, and K. S. Wagonei. C. H. Smeltzer and M. S. Viteles were employed as consultants to the Board during the period this material was being developed.

CIVIL SERVICE TESTING

NUMBER AND NAME CHECKING

In each of the following groups, some of the pairs of names and numbers are exactly the same while others are different.

You are to check, on the connecting line, the pairs which are different and indicate on your ANSWER SHEET the total number of such pairs in each group.

EXAMPLE:

61. John Smith	_____	John Smith
Wm. C. Burns	_____	Wm. C. Burns
Thos. Doe and Co.	_____	Thos. Doe Co.
Dart Salt Corp.	_____	Dart Salt Corp.
Bryant, Mitchell	_____	Bryant, Mitchell

In the example above, three pairs are different so you are to make a heavy mark in space number 3 opposite question No. 61 on your ANSWER SHEET, thus:

1 2 3 4 5
61 | | | | |

Be sure you have marked No. 61 on your ANSWER SHEET, then go on to No. 62. Work as fast as you can without making mistakes.

62. Auto Service Shop	_____	Auto Service Shop
Alister & McAlester	_____	Alister & McAlester
Van & Van de Vyere	_____	Van & Van de Vyere
Raymond Protschold	_____	Raymond Protschold
Kassanitsch Storage	_____	Kassanitsch Storage

62

63. Paul A. Anderson	_____	Paul A. Anderson
W. G. Gorton	_____	W. G. Gorton

83. 17287	_____	17287
27355453231	_____	27355453731
829	_____	829
53728	_____	53728
3212	_____	3212

83

84. 736291	_____	736291
62622621	_____	62622621

ALPHABETIZING

Rearrange the names in each of the following groups in the order in which they would appear in an alphabetic file.

List the names in alphabetic order in the blanks at the right. The number in parentheses after each name must be kept with that name during the alphabetizing. When you have arranged the names in the blanks, indicate the alphabetized order by making a heavy mark on your ANSWER SHEET in the appropriately numbered space opposite the proper question number.

For example, when you alphabetize the names in the first group below, you will find that the name followed by the number 3 in parentheses belongs after No. 46. You will therefore make a heavy mark in space number 3 after question No. 46 on your ANSWER SHEET, thus:

1 2 3 4 5
46 | | | | |

Alfred Anthony	(1)
Estelle Anthony	(2)
A. L. Anthony	(3)
Emily M. Anthony	(4)
Miss Birdie Anthony	(5)

46.	<u>Anthony, A. L. (3)</u>
47.	_____
48.	_____
49.	_____
50.	_____

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

NUMBER, NAME, AND ARITHMETIC CHECKING

DO NOT OPEN THIS BOOKLET UNTIL GIVEN THE SIGNAL BY THE PROCTOR

Read the following directions very carefully:

The inside pages of this booklet contain three tests which will be timed. The first test consists of pairs of numbers and the second test of pairs of names. If the numbers or the names of a pair are exactly the same, make a heavy mark in space number 1 on your ANSWER SHEET beside the corresponding number of that pair. If they are different, make a heavy mark in space number 2 on your ANSWER SHEET.

The third test consists of simple arithmetic calculations. If a problem is correct, make a heavy mark in space number 1 on your ANSWER SHEET beside the corresponding number of that problem. If it is not correct, make a heavy mark in space number 2.

SAMPLES DONE CORRECTLY

Question Booklet	Answer Sheet
1. 453-----435	1 1
2. 6125-----6125	2 1
3. William Johnson-----William Johnson	3 1
4. Abraham and Link Co.-----Abraham and Lind Co.	4 1
5. $4 + 8 = 12$	5 1
6. $3 \times 6 = 15$	6 1

NOW DO THE SAMPLES BELOW

7. 326-----326	7 1
8. 7418-----7814	8 2
9. Samuel Dillon-----Samual Dillon	9 2
10. Markwell and Gordon-----Markowell and Gordon	10 2
11. $14 \times 5 = 70$	11 2
12. $12 \div 6 = 3$	12 2

Whenever the proctor says "Stop," STOP WORK and listen carefully for further instructions.

FAILURE TO FOLLOW INSTRUCTIONS EXACTLY MAY LOWER YOUR SCORE IN THE EXAMINATION.

DO NOT OPEN THIS BOOKLET UNTIL GIVEN THE SIGNAL BY THE PROCTOR.

Make a mark in space number 1 if the numbers are exactly the same.
Make a mark in space number 2 if the numbers are different.

1. 616-----626
2. 4572-----4752
3. 0618-----0618

Make a mark in space number 1 if the names are exactly the same.
Make a mark in space number 2 if the names are different.

1. John L. Frankson-----John L. Frankson
2. Overholt Tobacco Co.-----Overholt Tobacco Co.
3. Time Recording, Inc.-----Time Recording, Inc.

Make a mark in space number 1 if the calculation is correct.
Make a mark in space number 2 if the calculation is incorrect.

1. $13 + 5 = 18$
2. $9 + 6 = 15$

CIVIL SERVICE TESTING

ALPHABETIZING AND SORTING

Each name below represents an addressed letter. You are to determine the number of letters addressed to each person. First, write the names of the Junior Visitors, Senior Visitors and the Interviewers in alphabetical order in the spaces provided. Then tabulate on the spaces provided the letters each received. Make a heavy mark on the ANSWER SHEET to indicate this number. For example, if the person received 3 letters make a heavy mark in space 3 beside the number of that name on the ANSWER SHEET. Thus, since Juanita Bates is the name which will be first when the Junior Visitors are arranged in alphabetical order, this name is written at the top of the Junior Visitor list. Tallying will show that she received two letters. Hence you will make a heavy mark in space number 2 beside Question No. 61 on the ANSWER SHEET.

Name	Position		Junior Visitor	Space for Tallying
Alberta Cummins	Junior Visitor			
Alfreda Swift	Junior Visitor			
Helen Cushman	Senior Visitor			
Rita Bauman	Junior Visitor			
Alberta Swift	Senior Visitor	61.		
Mary Petrey	Senior Visitor			
Eleanor Petrey	Junior Visitor	62.		
Rose Bowman	Senior Visitor			
Alberta Swift	Senior Visitor	63.		
Juanita Bates	Junior Visitor			
Jenny Betts	Senior Visitor	64.		
Jeanne Bolton	Interviewer			
Alberta Cummins	Junior Visitor	65.		
Mary Beck	Interviewer			
Mary Petrey	Senior Visitor			
Nathryn Snow	Interviewer			
Rayman	Junior Visitor			
	Senior Visitor			

SORTING

This is a sorting test. The city in which each person lives is represented by a code symbol. Determine the number of persons living in each city.

Use the blanks provided in the code list for purposes of counting. If the code symbol for a city is not listed, count it as Miscellaneous (No. 30). When you have finished sorting the names, count the number living in each city and indicate the total for each city by making a heavy mark in the proper space on your ANSWER SHEET.

For example, if it is found that four persons live in a city whose code symbol is RD-4, you would make a mark in space number 4 beside the question number of that code symbol, thus:

CODE LIST

APPLICANT'S NAME	CODE FOR CITY	APPLICANT'S NAME	CODE FOR CITY	
Moore	RB-8	Gray	JR-3	16. AB-3
Evans	SS-4	Cooper	FS-1	17. AK-4
Poster	MB-6	Force	RB-7	18. ER-5
Brown	HD-7	Burt	PT-2	
Jones	PT-2	Lewis	RB-8	
	AK-4			
Crown	VB-8	Murphy	WY-5	23. RD-4
Wells	MB-6	Kahn	VB-8	24. PT-2
Fink	KL-9	Rolfe	QR-1	25. QR-1
Sporn	QR-1	Borg	KL-9	
Call	WY-5	Hansen	FS-1	26. RB-8
Swift	SS-4	Rees	RB-8	27. SS-4
Harris	MB-6	Roth	UM-7	
Moses	ER-5	Prince	MB-6	28. VB-8
Brand	FS-1	Creel	SS-4	29. WY-5
Moon	AK-4	Yule	PT-2	30. Miscellaneous

SPELLING

51. It is a pleasure, unparalleled in my experience to recommend 61
62. John Gains, who has been my factory superintendant for ten years. 62
63. His imperturbable good nature and ineredable energy have made him 63
64. my most valued associate. Independent in his views, he is 64
65. ~~commodating to the needs of the situation. As~~ 65

Below is a list of the names and addresses on a temporary pay-roll, grouped according to the city in which each employee lives and arranged in order of the amount owed. Some of these names appear on the following page arranged in alphabetic order. You are to compare each name, street address, salary and amount of salary check with the same information listed on the next page. Pay no attention to names which are omitted from the table.

On the next page, check the names, street addresses, cities, and salaries which are not exactly the same as those on this page. Then make a heavy mark in the proper space on the ANSWER SHEET. The same marks indicate the total number of errors in each employee's record, thus: (1) one; (2) two; (3) three; (4) four; (5) none. Since but one error is counted for each name, street address, city, or salary in which errors occur, there can be, at the maximum, four errors in a record.

<u>Name</u>	<u>Salary</u>	<u>Name</u>	<u>Salary</u>	<u>Name</u>	<u>Salary</u>
<u>Albany</u>		<u>Cattaraugus</u>		<u>Chemung</u>	
Peter Metzer	\$99.90	Thomas B. Jarboe	\$227.40	Erman Celler	\$500.00
3626 Belmont Place		1316 East Master St.		2713 Dixmuth Ave.	
M. P. Iolas	76.90	John Frederick Hider	91.94	A. D. Washington	98.05
1219 Perry Ave. N.		108 4th St. S.W.		2934 Carlton Ave. S.	
Alfred Robertson	52.85	E. Edward Edwards	84.50	Petroff Bialstein	22.00
711 4th Ave. S.E.		1972 Kennedy Road		1237 Goddard Road	
		Pierot Maupassant	60.25	Montaigne Nettleton	8.45
<u>Kinghamton</u>	\$52.70	1832 20th Ave. N.E.		1204 Laidlaw	

EXAMPLE. In question number 151 you will find that there is an error in the name (the first initial should be W), an error in the street address (the house number should be 3119), and an error in the city (it should be Delhi). A check has been made in each of the proper columns at the right of the table to indicate these errors. You will, therefore, make a heavy mark in space number 5 after question number 151 on your ANSWER SHEET to show that there are 3 errors in the record on the first line of the list.

EMPLOYEE	STREET ADDRESS	CITY	NAME	STREET	CLAY	SALARY
151. Anber, L. Paul	6119 Hartford St. N. W.	Washington	✓	✓		151
152. Becheator, Rosemary	413 East Ninth St.	Canastota			✓	152
153. Breston, Petroff	1237 Goddard Road	Canastota				153
154. Ciesep, Anacovic	1981 Dale St. S. W.	Killbuckville				154
155. Ciesep, J. S.	2732 Wood Av.	Washington				155
						27.65

CIVIL SERVICE TESTING

PROOF READING

Reproduced below is a correct copy of a page of printing. On the right-hand page is a copy of the same material containing all kinds of errors. Count the number of errors in each line of the copy on the right-hand page. Make a heavy mark on the ANSWER SHEET in the space having the same number as there are errors in the line, thus (1) one error, (2) two errors, (3) three errors, (4) four errors, (5) no errors. Every deviation from the copy below (except length of line) is to be counted as an error. Only one error is to be counted for each word or number group in which errors occur. If you check each error as you come to it, you will find it easier to count accurately the number in each line.

The total number of applications for public assistance received during the last quarter of 1938 was 194,743. This represents an increase of 6.1 percent over the total for the previous quarter and a rise of 7.9 percent as compared with October-December 1937.

Home relief applications received, numbering 177,102, were 4.9 percent above the figure for the preceding quarter and 3.6 percent above the figure for the preceding year.

106.	The total number of applications for public assistance received	106
107.	during the last quarter of 1938 was 194,743. This represents	107
108.	an increase of 6.1 percent over the total for the previous quarter	108
109.	and a rise of 3.7 percent as compared with November-December 1937.	109
110.	Home relief applications received, numbering 177,102, were	110
111.	4.9 percent above the figure for the preceding quarter and 3.6	111
112.	percent above the number received in the last three months of	112
113.	1937. nearly 9.1 percent of all applications filled were account-	113

PARAGRAPH MEANING

DIRECTIONS FOR ANSWERING QUESTIONS 121 TO 146 INCLUSIVE:
Each paragraph below includes one word which spoils the meaning of the paragraph. This incorrect word is one of the five words which have numbers printed just above them. When you have found the incorrect word, make a heavy mark on your ANSWER SHEET in the space having the same number as the incorrect word.

121. Our present defeat by the machinery around us is a permanent thing, a plateau in our progress to a slaveless world. 121
122. The development of the higher organisms may be regarded as due to the coming together of these cells to form a

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

FOLLOWING DIRECTIONS

This is a test of your ability to follow directions.

You are to classify the employees of a department according to the salary each receives. The schedule of classifications is as follows

Class 1 \$1000 to 1099,
Class 2: 1100 to 1199,
Class 3 1200 to 1299,
Class 4 1300 to 1399,
Class 5: below \$1000 or above \$1400.

The salary for an account clerk is \$1100.
The salary for a mail clerk is \$1025.
The salary for a stenographer is \$1175.
The salary for a secretary is \$1350.

A Junior employee in any of these positions receives \$50 less than the amount shown, while any Senior employee receives \$100 more. For example, a Senior Stenographer receives \$1275 (\$100 more than a Stenographer) while a Junior Stenographer receives \$1125 (\$50 less than a Stenographer).

After reading the directions you are to classify the positions listed below according to the salary each receives, with these exceptions

1. Positions not mentioned in the above directions are to be placed in class number 5.
2. Individuals having five or more years' experience are to be placed in class number 5.
3. Individuals with less than 2 years' experience are to be placed in class number 5.

Indicate your answer by making a heavy mark on the ANSWER SHEET in the space having the same number as the salary classification.

EXAMPLES

A Stenographer with 3 years of experience will fall into Class 2, so a heavy mark is made in space number 2 on your ANSWER SHEET, thus,

1	2	3	4	5

A Senior Stenographer with 7 years of experience will fall into Class 5, so a heavy mark is made in space number 5 on your ANSWER SHEET, thus,

1	2	3	4	5

Question Number	Departmental Division	Position	Years Experience	Question Number
91.	Accounting	Account Clerk	4	91.
92.	Pay-roll	Stenographer	3	92.
93.	Administrative	Senior Stenographer	5	93.
94.	Filing	Junior Secretary	2	94.
95.	Clerical	Mail Clerk	7	95.
96.	Filing	Junior Account Clerk	5	96.
97.	Mailing	Stenographer	6	97.
98.	Clerical	Senior Account Clerk		98.

CIVIL SERVICE TESTING

T-F PAIRS IN FIVE-CHOICE FORM

This part of the examination consists of 50 questions, each made up of two statements, labelled A and B. You are to determine the truth or falsity of each of the statements. Having done so, you are to indicate your answers on the ANSWER SHEET as follows

1. If you consider that the answer to either or both statements in any question cannot be known, make a heavy mark in space number 1 on your ANSWER SHEET.
2. If you consider both statements in any question to be true, make a heavy mark in space number 2 on your ANSWER SHEET.
3. If you consider the first statement to be true and the second statement to be false, make a heavy mark in space number 3 on your ANSWER SHEET.
4. If you consider both statements to be false, make a heavy mark in space number 4 on your ANSWER SHEET.
5. If you consider the first statement to be false, and the second statement to be true, make a heavy mark in space number 5 on your ANSWER SHEET.

For your convenience, these directions are summarized below.

Mark 1, if none of the answers below applies.

Mark 2, if both statements are true.

Mark 3, if first statement is true, second is false.

Mark 4, if both statements are false

Mark 5, if first statement is false, second is true.

EXAMPLE

- (A) All public agency employees are happy.
 (B) Federal grants to States for public assistance will be drastically changed by 1950.

In the above question, the first statement is false. However, nobody can know whether the second statement is true or false. Consequently, the answer would be marked 1 as shown below.

1	2	3	4	5

76. (A) A person who does not look you in the eye is likely to be dishonest.
 (B) An interview with an emotionally upset client should be postponed until another day.

77. (A) Public records and documents are an optional source for verification of eligibility of applicants for public assistance by the visitor.
 (B) Information is not included in the index of a social service exchange as to the treatment given to a registered individual.

78. (A) All property of a recipient of old age assistance is considered part of the recipient's estate in the Probate Court.
 (B) The fact that an aged person is a recipient of a pension from some firm in industry is not an eligibility factor for

76

77

78

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TOPICAL FILING

The five classifications in a subject file are as follows:

1. Accounting (includes accounts)
2. Administration (includes personnel)
3. Maintenance (includes equipment and supplies)
4. Sales
5. Transportation

Below is a series of topical sentences, names of catalogs, and the like which you are to classify according to the above five divisions. If the statement would be most logically classified under accounting, make a heavy mark on the ANSWER SHEET in space number 1; if it refers to administration, make the mark in space number 2, and so on.

EXAMPLES

Make a mark in
space number

A copy of the building code	3
Regulations governing the wrapping of packages	5
A copy of corporation laws	2
Instructions to salesmen	4
Profit and loss statement	1

- | | |
|-----------------------------------------------------------------|-----|
| 121. There is no express office in Fraser. | 121 |
| 122. I am interested in learning more about your product. | 122 |
| 123. I fear the chief account clerk will have to be discharged. | 123 |
| 124. A list of freight stations. | 124 |
| 125. A | |

PUNCTUATION

The following selection contains errors in punctuation. Read each sentence through first to get its meaning. Then correct the errors by crossing out needless punctuation, changing incorrect punctuation, and supplying omitted punctuation. Consider each of the following punctuation marks as one error.

period .	semicolon	quotation mark
comma ,	colon	parenthesis) or (
hyphen -	apostrophe	

When you have made all necessary corrections in punctuation, count the number of errors which occur in each line and make a heavy mark in the appropriate space on the ANSWER SHEET, thus:

Space number 1 if there is 1 error
Space number 2 if there are 2 or more errors
Space number 3 if there are no errors

- | | |
|----------------------------------------------------------------------|-----|
| 151. Copyright laws have been in effect in the U.S. for more than | 151 |
| 152. one hundred years, the first statute being passed May 31, 1790. | 152 |
| 153. An author or owner, of unpublished material has a common-law | 153 |
| 154. The | |

CIVIL SERVICE TESTING

fangled ideas as scoring machines. Also, the State Technical Advisory Service of the Social Security Board was sufficiently interested in the mechanization of the selection process to encourage experimentation in that direction. Thus it happened that the Employment Board of the Pennsylvania Department of Public Assistance, a merit system agency whose history dates back only to the end of 1937, decided to score its examinations by machine.³ Since then, the number of agencies with scoring machine installations has been steadily increasing.

Scoring Civil Service Tests with the I.B.M. Machine: Procedures Ensuring Necessary Accuracy

Insofar as the early reluctance to adopt machine scoring was based on skepticism concerning its accuracy, it was on firm ground, for when a civil service agency puts a score on a test paper, that score must be accurate. There is probably nothing more likely to undermine the prestige and public acceptance, if not the very existence under law, of a civil service commission than the frequent, or even infrequent, discovery of errors in its work. To the public, the exact nature of the procedures used by a commission in scoring its papers are relatively unimportant so long as they are honest and produce *correct* results.

In the early days of scoring civil service tests by machine it was thought that the procedures which were satisfactory for scoring educational achievement and similar tests would be equally satisfactory for scoring civil service examinations, if certain additional precautions to spot errors were taken. Such procedures have, however, been abandoned in almost every instance in favor of a system designed specifically to meet civil service commission requirements of accuracy. The procedure now in use by the majority of agencies gives results that are probably as accurate as it is possible to obtain while human beings operate the machines and practical considerations render it absurd to recheck other than borderline scores beyond

³ Actually, the scoring was performed for the Employment Board by the Educational Records Bureau of New York City.

the point of finding more than one or two errors among thousands of scores.

This scoring system is suitable when the number of items answered correctly constitutes the written raw score, and depends for its accuracy upon recognition of the fact, as stated in the I.B.M. manual, that "the only truly accurate method of scoring is the one which takes into account every mark on the answer sheet which is intended as an answer, making allowance for questions answered more than once, and eliminating from the final score all stray marks not intended as answers by the examinee" (11). The five steps to this procedure are as follows:

1. Scanning papers for omissions and for items to which more than one answer has been indicated, and for the purpose of segregating sheets so poorly marked that they must be scored manually. A check mark is placed beside each omitted item and the number of items omitted is indicated in the box provided in the margin of the answer sheet. A horizontal line is drawn through each item answered more than once and the number of "surplus" answers indicated in the margin. (All marks on the answer sheet with the exception of those made by the examinee should, of course, be recorded with colored pencil.)

2. Scoring for rights on the machine.

3. Scoring for wrongs on the machine.

4. Totaling rights, wrongs, and omits (compensating for items answered more than once) and checking the total to see whether every item has been accounted for. If the total checks at this point, no further operations are necessary except manually scoring every 25th or 50th paper to provide a spot check of accuracy. (The additional precaution may be taken of manually scoring the answer sheets of all examinees whose scores range from two points below to one point above the passing point.)

5. Adjusting papers on which the total does not check in Step 4. When this is necessary, the paper goes to an adjuster whose job it is to determine the reason for the discrepancy and to correct it. Answer sheets which require such adjustment are then checked by a second adjuster to ensure accuracy. Answer sheets rejected in Step 1 as unsuitable for machine scoring are scored manually and checked by these adjusters or by others especially designated to perform this operation.

By means of the commoning key now available to scoring machine users, it is possible to perform a very useful screening operation in connection with Step 1, described above. This key may be inserted in the scoring rack between the sensing

CIVIL SERVICE TESTING

and resistance units and the machine then adjusted so that the meter will indicate, for any given side of an answer sheet, the number of items attempted. For papers for which the machine adjusted in this way indicates "no omits," the task of scanning may be reduced to looking for items with multiple markings and papers which it is desirable to score manually.

A further refinement to this system may be wired into the machine so that the meter, instead of indicating the number of items attempted, will read the number omitted. This is accomplished by adjusting the circuit so that sufficient current will flow through the meter initially to indicate the total number of items in the test. Then, when an answer sheet is placed in the machine, the number of items attempted will automatically be subtracted from the initial reading, causing the meter to indicate the number of omissions.

Before leaving the subject of scoring, a few words of caution may be in order. The machine process in its present state of perfection is a big improvement over most other scoring techniques at present available for use by civil service jurisdictions. It has not, however, reached the state of refinement where it can be taken completely for granted that nothing will go wrong after the machine has been set up. Since every once in a while something does go wrong, it is necessary, to avoid later grief, not only to set up checks and controls but to require strict adherence to them on the part of all staff members charged with any scoring responsibility.

Other Machine Methods of Scoring

The I.B.M. scoring machine seems to offer, for ordinary civil service use, what appears to be the best all-around solution to many of the most annoying scoring problems confronting the medium or large size civil service agency. In addition, as will be noted later, this particular machine may be adapted for use in connection with several other examining tasks, including certain research projects that every civil service agency has in mind carrying through just as soon as the staff and the time are available.

There are, however, at least two additional scoring approaches classifiable under the head of "mechanical" that have been used to some extent by agencies conducting large numbers of examinations.

The first of these makes only partial use of a mechanical device—in this case a multilith, mimeograph or printing press—and is more accurately a technique for facilitating manual scoring than a procedure for scoring by machine. Separate answer sheets are used on which the examinee indicates his answers to multiple-choice or true-false items by checking appropriately numbered spaces. A multilith or mimeograph stencil is then prepared so that, when the answer sheets are run through the duplicating machine, a line connecting the correct answers will be printed over the response positions. When this has been done, it is possible for a scoring clerk to determine the number of correct answers simply by counting the number of responses marked in positions coinciding with the overprinted line.

This combination machine-manual method is considerably more rapid and accurate than manual scoring accomplished by placing a key alongside or over the answer sheet. By altering the procedure slightly it may also be used to advantage with completion-type questions. It requires, however, a skilled duplicating machine operator and the use of a duplicating machine capable of very accurate registration and extremely little spoilage. The multilith satisfies these requirements particularly well, and printing may, of course, be employed. On the other hand, the mimeograph appears to be less satisfactory.

A procedure, described by Iffert, Bloom, and Beum (6), for scoring multiple-choice tests by means of tabulating machines has apparently also been used with some success, although its chief value would appear to be in connection with the conduct of examining programs that are quite intimately tied in with research projects. When this particular method is employed, the examinee's answers are usually placed directly in the test booklet from which they are later punched into Hollerith cards and scored by successive runs through a sorter.

Once these cards are punched they are also available for research purposes and it is comparatively easy to conduct item analyses and compute correlations with them.

Scoring Graphic Rating Scales by Machine

Graphic rating scales may be, and have been, scored by the I.B.M. scoring machine. "By using the aggregate weighting unit of the machine it is possible to obtain the aggregate weighted average of as many as 30 variables, each varying in size from 1 to 100 and (in groups of three) weighted from 0 to 20" (10).

In utilizing this feature of the machine the rating scale is usually designed so that it is necessary to draw a horizontal line for each characteristic rated. The length of each such line determines the score for that particular characteristic, and the weighted total for all characteristics is indicated on the meter in the same fashion as any other score. The horizontal lines should, of course, be drawn with a special pencil, and may be made by the rater himself or be drawn in later by a clerk.⁴ When the latter plan is used, the rater checks (with a colored pencil) the point on each line that represents the rating he wishes to assign and the clerk simply draws a line from the origin (left) to each check mark. Graphic scales of this type may be used in connection with oral interviews, service ratings, or performance test ratings.

Scoring Training and Experience by Machine

Many civil service commissions include a quantitative rating of training and experience in the test battery for a majority of the classes of positions for which they conduct examinations. It now appears quite likely that a considerable portion of the computational work connected with the use of the type of training and experience rating scale (14) employed, in one

⁴ In several informal studies conducted by an agency which formerly used this type of rating sheet in large quantities it was found that where raters made check marks only, it was apparently faster to score scales of this kind manually (by having a clerk place a stencil over the rating sheet and add the numbers on a comptometer) than to go to the trouble of drawing a line for each characteristic before running the sheets through the scoring machine.

form or another, by numerous agencies throughout the country may be performed by machine.

This possibility was brought closer to realization recently with the development⁵ of a tentative form of machine-scorable training and experience rating scale which, while it still remains to be tried in an actual test situation, looks very much as though it will not only work but possibly be at least a partial or preliminary answer to the mechanization of this phase of the selection process.

This adaptation of the machine makes use of the principle of the previously mentioned commoning key. In marking the scale the rater simply blackens each space that corresponds to a type of training or experience possessed by the examinee whose application is under consideration. As an example of the possibilities of this approach, provision has been made in the initially developed form of the scale for recognition of up to 15 years—in six-month steps—of each of four levels of related experience, three levels of related under-graduate and graduate study, and the possession of academic degrees.

Scoring Service Ratings by Machine

Present service rating instruments take various forms, many of which may be scored by machine. Before deciding to adopt such a procedure, however, the numerous factors involved should be given careful consideration, and the decision based upon the extent to which machine scoring will contribute to the economy, speed, and all-around efficiency with which this particular phase of the program may be administered.

Service rating scales of the graphic type may be scored by setting them up to utilize the aggregate weighting feature of the machine. In addition various adaptations of the graphic approach may be employed which, depending upon the particular situation at hand, may be scored by using either the ordinary answer key form alone or in conjunction with the more recently available commoning key, with the latter set-up offering considerably the greater possibilities. Service rating

⁵ This machine-scorable scale was the outgrowth of a discussion participated in by E. C. Schroedel, G. C. Slougher, J. H. Pockrass, and S. W. Koran.

forms of the check-list variety may also rather easily be adapted to scoring by machine.

Item Analysis With the Scoring Machine

A recently developed attachment to the scoring machine is the graphic item counter, which is available as optional equipment. This device consists of a plugboard having a plugging position for each of the 750 response positions on the standard answer sheet and for each of 90 counters. By means of plug-wires, any response position may be connected to any counter. When the appropriate response positions and counters have been wired together, the plugboard is inserted into the machine in the position normally occupied by the scoring rack.

Using this attachment it is possible to secure, in a single run through the machine, a graphic count of the marks placed in up to 90 response positions on 100 answer sheets. If more than 100 sheets are involved in the study, a separate graphic count must be made for each group of 100 sheets. If more than 90 response positions are to be analyzed in a given test, the plugboard must be rewired and the sheets run through again for the additional responses. Thus, if in a given analysis, it is desired to determine the number of individuals who correctly answered each of the 150 five-choice items on a single side of an answer sheet and the population of the study is 175, it is necessary to run the 175 sheets through the machine twice, making separate graphic counts of the first 100 and the last 75 on each of the two runs. Items 1 to 90, inclusive, may be analyzed on the first run, and the remaining 60 items (91 to 150, inclusive) on the second run.

If the item analysis is of the variety that requires information concerning the examinees' selection of each of the five possible responses to the 150 items, the sheets will have to be run through the machine nine times to obtain this information for each of the 750 possible responses. Whether the items will be analyzed to the extent of determining the number of examinees selecting each possible response or be confined to determining the number of examinees selecting the correct answer

will, of course, depend upon the use or uses for which the data are intended. The speed of operation of the machine equipped with the item analysis unit has been reported as ranging from 400 to 500 papers per hour when 90 responses are analyzed on each paper. This is considerably faster than any clerk can perform the job manually.

Computing Reliability Coefficients, Standard Error of Measurement and Intercorrelations with the Scoring Machine

During the past few years several techniques have been developed for using the test scoring machine to facilitate the computation of such useful measures as intercorrelations, reliability coefficients, and the standard error of measurement. While it is beyond the scope of this presentation to go into the derivation of the formulas that have been developed or to describe the procedures at any length, mention will be made of a few of the more important of these applications in order to illustrate the variety of the scoring machine's uses in connection with examination research.

One kind of research, that of investigation into the validity of individual items by means of the technique of item analysis, has already been mentioned. Hoyt (5) recently described a method of computing test reliability which makes further use of some of the data obtained when such an item analysis is performed. The procedure which Hoyt suggests was developed as a practical and simplified application of Richardson and Kuder's (9, 15) "method of rational equivalence," which produces a coefficient of reliability that in certain respects appears to be superior to that obtained by using the split-half correlation method with the Spearman-Brown formula. No data beyond those obtained when a test is scored and an item-analysis performed are required for substitution into the following formula (5):

$$r_{tt} = \frac{n}{n - 1} \cdot \frac{kS_s + S_i - T(T + k)}{kS_s - T^2}$$

in which r_{tt} is the reliability of the test, n is the number of items in the test, k is the number of subjects taking the test,

T is the sum of the scores obtained by all the subjects, S_s is the sum of the squares of the scores obtained by the subjects, and S_i is the sum of the squares of the number of correct responses to each item.

Although the Kuder-Richardson technique has numerous advantages which make its wider adoption quite likely, many investigators in the field of civil service examinations may have to continue to obtain most of their reliability coefficients by means of the split-half method used in conjunction with the Spearman-Brown formula. This, at any rate, will probably continue to be the situation unless item analysis data are available for substitution into a formula such as the one above or the simpler formula presented by Kuder and Richardson⁶ is not appropriate in the specific situation.

When it is known at the outset of scoring a given test that a split-half reliability coefficient will be required, it is possible to prepare the scoring matrices so that separate scores for the odd-numbered and for the even-numbered items will be obtained when the answer sheets are run through the machine for the first time. This may be accomplished by keying the odd-numbered items in the usual fashion (that is, as rights), and the even-numbered items as wrongs (that is, preparing the scoring matrix so that even-numbered items answered correctly will be indicated when the selector switch is in the wrongs position). To secure the total rights score when the machine is set up in this way, the selector switch need only be moved to the $R + W$ position. When the selector switch is moved to the R position the meter will indicate the number of odd-numbered items answered correctly, and when it is in the W position, the number of even-numbered items answered correctly.

Far from being extra work, this procedure possesses the

⁶ It should be noted that Kuder and Richardson (9) have derived a simpler formula (No. 21 in the article referred to) which can be computed in two or three minutes, given the number of items in the test, the average score, and the standard deviation of the scores. It gives a slight underestimate of the true reliability of a test. For the more reliable tests the estimates obtained by this formula are usually from .01 to .03 less than those obtained by use of the more rigorous formulas presented.

advantage of providing a useful check on the total score. As only half of the sets of contacts are connected to the meter when the switch is in either the R position or the W position, the effect of stray marks on the reading is frequently minimized to the point where, on some papers, the total of the separate R and W readings (added without the use of the scoring machine) may provide a more accurate score than the $R + W$ (total rights) reading taken alone. The reason for this is, of course, that while the effect of stray marks and poor erasures may be sufficient to influence the score when 150 sets of contacts are in the circuit, their effect may not be noticeable when only 75 sets of contacts are involved at a time.⁷

It might be well, at this point, to call attention to the fact that when the "rights plus wrongs plus omits" scoring procedure described earlier in this paper is employed with a single machine set up to read rights and wrongs on a single insertion, it is not possible to secure the split-half scores at the same time in the fashion just suggested. This problem may, however, be solved by reading the rights on one run through the machine and the wrongs on a subsequent run.

Several formulas are available for determining split-half reliability. In place of the orthodox product-moment formula, some investigators prefer, under certain circumstances, a formula which requires the substitution of values obtained from the odd scores and total right scores only. As described by Mosier (13) this formula takes the form:

$$r_{os} = \frac{r_{ot}\sigma_t - \sigma_o}{\sqrt{\sigma_t^2 + \sigma_o^2 - 2r_{ot}\sigma_o\sigma_t}}.$$

⁷ The problem presented by the necessity for manually scoring considerable numbers of answer sheets because the effect of stray marks causes the total of rights, wrongs, and omits to differ from the total number of items is worthy of attention. The kind of solution suggested above in connection with the split-half scoring of rights may, if necessary, be adopted as a regular practice in scoring wrongs as well as rights. Because of the considerably larger answer sheet area exposed to "live" contacts when the wrongs score is being determined, such readings are usually affected by stray marks and poor erasures to a greater degree than rights scores. When split-half scores are not required, the answer sheet area may, of course, be divided into two or three sections by punching appropriate field selection holes so that a separate reading may be obtained for each area.

Still another approach—and one which has been gaining considerable favor recently with scoring machine users—makes use of the fact that once the machine has been set up to indicate odds and evens, the difference between these values may be obtained by simply turning the selector switch to the R-W position. The standard deviation of the difference scores thus obtained for a given test will, as has been pointed out by Rulon (16), equal the standard error of measurement of the scores in that group. A split-half reliability coefficient may then be obtained by substituting in the following simple formula (2):

$$r_{tt} = 1 - \frac{\sigma_o^2}{\sigma_w^2}$$

in which r_{tt} is the reliability of the test, σ_o is the standard deviation of the difference between the odd and even scores, and σ_w is the standard deviation of the ordinary test scores including both odds and evens.

A method making it possible to use the scoring machine for calculating tables of intercorrelations has been developed by Kuder (8) but has apparently not been very widely used, perhaps because of the laborious clerical work involved in preparing the coded answer sheets required for the computations. The work thus involved has lately been reduced, however, in a revised and simplified procedure suitable for studies involving up to 150 cases. The Kuder technique is an adaptation of the Royer-Toops method of obtaining correlations from Hollerith cards on which geometric codes of scores for each variable have been punched. Kuder's approach has been to use an answer sheet and stencil for each code, but he does not recommend substituting the scoring machine for Hollerith equipment when the latter is easily available (8).

Kuder has also pointed out that the scoring machine is excellently suited for obtaining tetrachoric coefficients of correlation and that the procedure for doing so is relatively simple. Since in tetrachoric correlation each variable is divided into a dichotomy instead of into class intervals, the amount of "coding" and clerical work involved is reduced to a minimum.

Tabulating Equipment

It has been possible to note, during the past few years, a considerable increase in the number of civil service agencies that have adopted mechanical procedures for handling, in addition to scoring, such other operations as assigning candidates to the written, oral, and performance tests; computing grades; notifying candidates of their eligibility and ineligibility; establishing registers of eligibles and lists of candidates who did not qualify; certifying names from registers; maintaining miscellaneous personnel and payroll records; and aiding in the conduct of research (3, 4). Let us examine briefly some of the ways in which Hollerith equipment may be used to facilitate the conduct of some of the operations connected with the processing of examinations, keeping in mind, however, the unlikelihood that a machine installation would prove particularly economical for conducting these specific operations to the exclusion of the numerous related tasks just enumerated.⁹

One of the more important jobs which it is possible to perform successfully with Hollerith cards is that of converting, weighting, and combining examiners' scores on the various components of the examination, transmuting the results into final grades, adding veterans' credit and other bonuses, and listing, in order of final grade, the names of those who have qualified.

The tabulating card used for this purpose includes fields which provide columns into which may be punched such data as the following, depending upon the needs of the given situation: identification and file numbers; class of position; written, training-experience, oral, performance, and service rating raw and converted (or weighted) scores; total converted score, final grade, veteran's credit, rank, *et cetera*. Written raw scores and identifying data are punched into what are called "detail" cards with the electric key punch and are verified by means of the mechanical verifier. These cards are then ar-

⁹ An exception to this might be the situation in which certain equipment of another agency is available for part-time use so that the only additional machines required are a key punch and verifier, and possibly a sorter.

CIVIL SERVICE TEST

ranged in order of identification number by means of the horizontal sorter. Raw scores on each subsequent part of the examination battery are usually first punched into "scratch" cards which, after being verified, are sorted according to identification number in the same fashion as the detail cards. When this has been done, the data on the scratch cards are transferred to appropriate columns of the detail cards through use of the automatic reproducing punch (3).

After the raw scores have been punched into the detail cards, it is usually necessary to convert them into whatever variety of transmuted score the agency uses for the purpose of assigning the announced weight to each component of the examination. The raw score data which ordinarily serve as the basis for computing the conversion tables are easily secured from the cards by sorting them by raw score and running them through a numeric tabulator.

Several methods of transferring the transmuted scores to the detail cards may be used: Conversion tables may be prepared for use by key punch operators who determine the converted score corresponding to each raw score and punch that figure into the detail card (4). A second method that may be employed calls for using the automatic multiplying punch for the purpose of multiplying the raw score by some constant (for example, the reciprocal of the number of questions in the written test, if a percentage is desired). A third method makes use of prepunched master cards each of which contains a possible raw score and its corresponding conversion. When using this arrangement, both the master cards and the detail cards are sorted by raw score and run through an automatic reproducing punch which transfers the converted scores from the master cards to the detail cards at a high rate of speed (3). When the transmuted scores for all components of the examination have been entered, they may be totalled and the sum punched into the appropriate columns of the detail card. If this total requires further conversion, the process involved is identical to that of transmuting individual raw scores and may be carried out in any one of the three ways mentioned.

In addition to serving the purposes for which they were designed, the punched cards used in arriving at the examinee's final grade and register position are also available for numerous research uses. Agencies to which the use of a Hollerith installation is available are in the fortunate position of being able to perform many of the research jobs discussed as possible with the scoring machine and, in addition, to make use of the amazing flexibility of the punched card method to conduct types of research which, because of the amount of clerical and statistical work sometimes involved, are all but impracticable when attempted without such aid.

REFERENCES

1. Dunlap, Jack W. "Problems Arising from the Use of a Separate Answer Sheet," *Journal of Psychology*, X (1940), 3-48
2. Flanagan, John C. "Note on Calculating the Standard Error of Measurement and Reliability Coefficients with the Test Scoring Machine," *Journal of Applied Psychology*, XXIII (1939), 529
3. Hawthorne, Joseph W and Morse, Muriel, *Business Machines in Public Personnel Administration*, Los Angeles: City Civil Service Commission, 1940, 43 pp.
4. Horchow, Reuben *Machines in Civil Service Recruitment* Chicago. Civil Service Assembly of the U. S. and Canada, Pamphlet No 14, 1939, 43 pp.
5. Hoyt, C J "Note on a Simplified Method of Computing Test Reliability," *Educational and Psychological Measurement*, I (1941), 93-95
6. Iffert, R. E., Bloom, B. S., and Beum, C. O. *Another Test-Scoring Procedure A Method of Scoring Short Tests on the Hollerith Sorter*. Columbus: Ohio College Association Bulletin, No. 118, Mimeographed, February 1940, 7 pp.
7. Koran, Sidney W "Adapting Tests to Machine Scoring," *Journal of Applied Psychology*, XXIII (1939), 709-719
8. Kuder, G Frederic. "Use of the International Scoring Machine for the Rapid Calculation of Tables of Intercorrelations," *Journal of Applied Psychology*, XXII (1938), 587-596.
9. Kuder, G. F., and Richardson, M. W. "The Theory of the Estimation of Test Reliability," *Psychometrika*, II (1937), 151-160
10. *Machine Method of Scoring and Analyzing Examinations*. New York: International Business Machines Corporation, undated, 14 pp
11. *Machine Methods of Test Scoring. Manual of Procedures*. New York. International Business Machines Corporation, 1940, 7 pp.
12. *Manual of Instruction for the International Test Scoring Machine*. New York: International Business Machines Corporation, 1939, 20 pp.
13. Mosier, Charles I. "A Short Cut in the Estimation of Split-Half Coefficients," *Educational and Psychological Measurement*, I (1941), 407-408.
14. Pockrass, Jack H "Rating Training and Experience in Merit System Selection," *Public Personnel Review*, II (1941), 211-222
15. Richardson, M. W., and Kuder, G. F. "The Calculation of Test Reliability Coefficients Based on the Method of Rational Equivalence," *Journal of Educational Psychology*, XL (1939), 681-687.
16. Rulon, Phillip J "A Simplified Procedure for Determining the Reliability of a Test by Split Halves," *Harvard Educational Review*, IX (1939), 99-103.

PREDICTIVE VALUE OF CERTAIN "LAW APTITUDE" TESTS¹

E. L. WELKER and T. W. HARRELL
University of Illinois

THIS PAPER REPORTS the second of a series of studies analyzing the abilities necessary for success in law school. An earlier study² showed that pre-law grades from one school, the University of Illinois, correlated higher with law grades than did the Ferson-Stoddard *Law Aptitude Examination*. Combining the test and pre-law grades did not significantly improve the prediction. The homogeneous parts of the law aptitude test were correlated separately with law grades and showed that the memory case questions gave an unquestionably insignificant correlation. This result is interesting since the memory material seems to represent a popular stereotype of what a law student has to do.

Several other investigators have reported comparisons between test scores and law-school grades, but apparently no one has previously reported a detailed attempt to analyze the relation between separate law course grades and part scores of "law aptitude tests." The ultimate aim of these studies is of course to discover tests that will lead to the more valid prediction of law school success.

The variables included in this study are listed in Table 1. It will be noted that the tests used were the homogeneous parts of the Ferson-Stoddard *Law Aptitude Examination*, in addition to the homogenous parts of other selected tests—the Yale *Legal Aptitude Test*, the American Council on Educa-

¹This study was made possible through the generous cooperation of Dean Albert J. Harno, University of Illinois College of Law.

²T. W. Harrell, "Predicting Success of Law School Students." *American Law School Review*, IX (1939), 290-292.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

tion *Psychological Examination*, and the comprehension and speed tests of the *Minnesota Reading Examination*. Some of the pre-law grades (variable 26) were approximated where students attended a school other than Illinois. Average law grades for the first semester as well as course grades for the five first-semester courses were used as criteria.

TABLE 1. NAMES AND DESCRIPTIONS OF VARIABLES

Variable Number	Description	Name
1	Interpretive case	Person-Stoddard Law Aptitude Exam Part 2-A
2	Completion	Person-Stoddard Law Aptitude Exam. Part 2-B
3	Relevant facts	Person-Stoddard Law Aptitude Exam. Part 2-C
4	Logical inferences	Person-Stoddard Law Aptitude Exam Part 3
5	Matching	Person-Stoddard Law Aptitude Exam. Part 4
6	Memory case	Person-Stoddard Law Aptitude Exam Part 1-B
7	Arithmetic case	ACE Psychological Examination (1938)
8	Pattern analogies	ACE Psychological Examination (1938)
10	Completion	ACE Psychological Examination (1938)
11	Artificial language	ACE Psychological Examination (1938)
11	Artificial language	ACE Psychological Examination (1938)
12	Same-opposite	ACE Psychological Examination (1938)
13	Reading speed	Minnesota Reading Examinations
14	Reading comprehension	Minnesota Reading Examinations
19	Word relations	Yale Legal Aptitude Test Group I
20	Opposites	Yale Legal Aptitude Test Group II
21	Word analogies	Yale Legal Aptitude Test Group III
22	Logical inferences	Yale Legal Aptitude Test Group IV
23	Memory case	Yale Legal Aptitude Test Group V
24	Interpretive case	Yale Legal Aptitude Test Group VI
25	Definitions	Yale Legal Aptitude Test Group VII
26	Pre-Law Grades, including those approximated from other schools	
27	Average First-Semester Law Grades, Univ. of Illinois College of Law	
28	Course Grades in Contracts	First Semester, Univ. of Illinois Col. of Law
29	Course Grades in Torts	First Semester, Univ. of Illinois Col. of Law
30	Course Grades in Remedies	First Semester, Univ. of Illinois Col. of Law
31	Course Grades in Criminal Law	First Semester, Univ. of Illinois Col. of Law
32	Course Grades in Possessory Estates	First Semester, Univ. of Illinois Col. of Law

The subjects were 133 male Law College freshmen at the University of Illinois. Seventy-eight of these entered in the fall of 1938 and 55 in the fall of 1939. The means of the two groups on both test scores and grades appeared similar enough to justify combining the data for the two years into one study.

The product moment coefficients of correlation between each of 21 test scores and average first-semester law grades are shown in Table 2. Insignificant correlations, i.e., those

"LAW APTITUDE" TESTS

less than .17, for which the chances that such a coefficient of correlation will occur in an uncorrelated population are more than 5 in 100, are omitted. Barely significant coefficients, i.e., those between .17 and .22, where the chances are more than 1 in 100, that such a coefficient will occur in an uncorrelated population, are in parentheses. No correction has been made for attenuation or the unreliability of the variables.

TABLE 2
PRODUCT MOMENT COEFFICIENTS OF CORRELATION WITH FIRST-SEMESTER LAW
GRADES N — 133

Variable Number	Correlation with First Semester Law Grades
1	—
2	—
3	(.17)
4	.28
5	.31
6	—
7	.23
8	(.17)
9	.24
10	.28
11	.31
12	—
13	—
14	.25
19	.25
20	.39
21	.33
22	.30
23	(.19)
24	—
25	—
26	.49

Note: Insignificant coefficients are omitted and barely significant ones are parenthesized.

In evaluating some of the parts of the *Yale Legal Aptitude Test*, it should be noted that the scores reported do not represent separate sections with individual time limits. The test is made up of three parts which are separately timed. The parts are not homogenous as to the type of item used. The first and third parts are composed of four item types: Word Relation, Word Opposites, Word Analogies, and Logical Inferences. These are arranged in cycle-omnibus form with 10 items of the same type together. The total number

of items of each type is 40. The second part is made up of three additional kinds of questions. First are 20 memory items dealing with a case that was presented at the beginning of the test—before Part 1. Next are 40 items of the Interpretive Case variety. Finally there are 20 Definition questions.

Mr. W. E. Kline of the Yale Personnel Bureau writes that the Interpretive Case and Definition items have consistently yielded only low correlations with grades. This result is explained by the fact that these types of questions do not appear early enough in a timed section for reliable scores to result. Consequently a new form of the Yale test is being put together. "It contains seven sub-tests, each of which is homogeneous and individually timed."

Tests correlating clearly significantly with law grades, as shown in Table 2, are, in order of their coefficients from high to low: Yale Opposites, Yale Word Analogies, Ferson-Stoddard Matching, ACE Artificial Language, Yale Logical Inferences, Ferson-Stoddard Logical Inferences, ACE Completion, Minnesota Paragraph Reading Comprehension, Yale Word Relations, ACE Number Series, ACE Arithmetic Tests adjacent in order seldom if ever have coefficients that are significantly different.

It is recognized that a completely thorough understanding of the interrelations of the variables calls for a factor analysis. Such a study is planned. All intercorrelations have been computed.

Tests which correlated barely significantly with law grades, as shown in Table 2, are: Yale Memory Case, Ferson-Stoddard Relevant Facts, and ACE Pattern Analogies.

The following tests did not correlate significantly with the first-semester mean: Ferson-Stoddard Interpretive Case, Ferson-Stoddard Analogous Case, Ferson-Stoddard Memory Case, ACE Same-Opposite, Minnesota Reading Speed, Yale Interpretive Case, and Yale Definitions.

None of the correlations is as high as .40. The Yale test correlates slightly higher than any other test total. Some of

"LAW APTITUDE" TESTS

the American Council sub-tests and the Minnesota Reading Comprehension correlate significantly, while some of the so-called law aptitude sub-tests do not.

It was mentioned above that the previous study showed that the memory case questions in the Ferson-Stoddard test correlated insignificantly with law grades. This result is confirmed here, but the Memory Case in the Yale examination does give a barely significant correlation. Mr. Kline writes that the memory questions correlated .33 with first-year grades of the Yale Law freshmen of 1940

Pre-law grades correlated .49 with first-semester law grades. This coefficient is higher than any with test scores, but considerably lower than that reported in the previous paper. One explanation for the lower coefficient is the lessened accuracy of the present pre-law grades. These include grades at schools other than Illinois plus those at Illinois. Previously only Illinois pre-law grades were included. Where grades from different schools are combined, it seems unlikely that the result will be as reliable a test of values as those from one school, due to differences in grading systems. Another reason for the decreased correlation between law grades and pre-law grades is that the present group is more homogeneous for pre-law grades. This situation was occasioned by raising the requirement for entrance for Illinois students having only 3 years' credits from a grade-point average of 3.0 to 3.25.

The product moment coefficients of correlation between each of five law grades and the 21 test variables are shown in Table 3. Again, insignificant coefficients have been omitted, and barely significant ones parenthesized as in Table 2. Variable 30, Remedies, correlated significantly with 12 test scores; variable 28, Contracts, with nine; variable 29, Torts, with nine; variable 32, Possessory Estates, with four; and variable 31, Criminal Law, with only two. Legal aptitude tests measure more nearly what is required to master Remedies, Torts, and Contracts, than they measure what is required to understand Criminal Law and Possessory Estates.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

The significance of these differences has not been tested. Part of the differences could be otherwise explained if the reliability of the grades varied markedly from one course to another, but their reliabilities are unknown. Some thought has been given to estimating the reliabilities using the Kuder-Richardson method, since it does not demand split-half scores. This has not been done because the number of items represented by the law grades is nonexistent. Some conversation with lawyers suggests that Remedies does demand more reasoning than does Criminal Law, which requires greater memorization.

TABLE 3.
PRODUCT MOMENT COEFFICIENTS OF CORRELATION BETWEEN EACH OF 5 LAW GRADES AND 21 TEST SCORES

Variable Number	28	29	30	31	32
1	—	—	—	—	—
2	—	—	—	—	—
3	—	(.17)	(.20)	—	(.19)
4	.32	.24	.31	—	(.22)
5	.28	.31	.34	(.19)	.25
6	—	—	—	—	—
7	.23	(.19)	.25	—	.24
8	—	(.22)	.23	—	—
9	(.22)	(.19)	.30	—	(.18)
10	.26	.29	.30	.26	(.17)
11	.37	.26	.41	(.20)	(.22)
12	—	—	(.18)	—	—
13	—	—	—	—	—
14	(.21)	.24	.26	—	—
19	(.21)	.24	.35	—	—
20	.39	.35	.47	.26	.30
21	.29	.32	.44	(.22)	.24
22	.25	.27	.39	(.22)	—
23	.25	(.22)	—	—	(.19)
24	—	—	(.21)	—	—
25	—	—	(.22)	—	—

Note: Insignificant coefficients are omitted and barely significant ones are parenthesized.

It will be noted that three of the correlations with Remedies are higher than any of those with the semester means. The differences are scarcely reliable

It can be tentatively concluded that while no legal aptitude test correlated as high with law grades as do pre-law grades, the most predictive tests are those that call for reasoning rather than memory. The reasoning tests may use

"LAW APTITUDE" TESTS

words or numbers for symbols, but there seems to be an advantage for the former, as might be expected.

Each of the two legal aptitude tests correlates higher with pre-law grades than with law grades. This difference might be explained if the pre-law grades are more reliable than the law grades. The authors have not been able to determine the reliability of either. The law grades might be expected to be more reliable from the fact that the law course is more homogeneous than is the varied pre-legal curriculum. On the other hand, grades based on 6 to 8 semesters of pre-law work, because of the increased reliability with additional length, would be expected to be more reliable than grades from a single semester of law.

Since the two so-called "legal aptitude" tests correlate lower with law grades than with other college grades and since several tests that are not called "legal aptitude" correlate higher with law grades than several that are putative measures of law-school success, the question is raised as to the possible existence of a factor or factors of legal aptitude. The factor analysis of these data may contribute a clearer answer to the question.

AN EXPLORATORY STUDY OF SOCIAL GUIDANCE AT THE COLLEGE LEVEL¹

MARGARET GLOCKLER ALDRICH

University of Minnesota

WITHIN THE LAST 10 years there has been an increasing emphasis on guidance at the college level. This movement is important, but it is significant that the personnel workers in institutions of higher education have been concerned almost exclusively with educational and vocational problems. In some cases, however, college authorities have come to realize that there are certain social problems and adjustments which should be considered. Many college programs fail to provide social stimulation and opportunity for participation. Extra-curricular activities have developed on college campuses to fill this need. Most of these activities have been developed on the basis of student initiative in spite of, rather than because of, faculty approval.

An interest in the development of social adjustment in colleges led to the following experimental evaluation of social guidance at the college level.

Naturally, the valuation of guidance has lagged far behind the development of guidance techniques. There have been several attempts to determine the effects of diagnosis and treatment of educational and vocational problems [Beaumont

¹This study was undertaken at the suggestion of Professor D. G. Paterson of the University of Minnesota. His advice and interest made its completion possible. Dean E. G. Williamson, Dr John Darley, and the counselors of the University of Minnesota Testing Bureau, as well as various extra-curricular organizations on that campus, made possible the execution of the problem.

(1), Wrenn (10), and Williamson and Bordin (8)]. These studies indicate the possibilities in this field. However, there has been no such study of social guidance, even though several workers have recognized the need for evaluating this type of guidance [Tuttle (7), Livingood (5)]. Others [Burke (3), Mallay (6), and Williamson and Darley (9)] have attempted to distinguish the socially "well" adjusted from the socially "poorly" adjusted. But in the psychological literature of the last five years there is no report of a study of the effect of any particular controlled factor on social adjustment

Several investigations at the University of Minnesota have demonstrated the need for further concern with the social and extra-curricular program [Chapin (4), Brown (2)]. The more extensive of these was that by Brown in 1934. She found that one third of the students spend no time or money in activity participation and concluded that "students most needing social contacts were those who profited least from the opportunities offered." (2:263)

These conclusions might be interpreted to mean that there is little hope of better adjusting asocial students. Another possibility is that if these students who did not participate were given an intensive well-directed program of participation, they might become better adjusted socially. Is it possible to lead the asocial to activities and find any changes in their social interests and attitudes?

The essential plan of this research was to expose students to certain social influences and measure any changes resulting from the contacts formed. To be of value, it was necessary that these influences be normal extra-curricular and counseling activities available to all college students. It was also necessary that the control group technique be used to determine what would occur without these special influences. This need led first to a consideration of possible methods for measuring changes that occur in the social adjustment of college students. To make the group as homogenous as possible, it seemed advisable to limit the study to freshman girls. Since all the

EXPLORATORY STUDY OF SOCIAL GUIDANCE

girls were to be treated as a part of a normal counseling program, it was further necessary to study only girls who had gone through the University Testing Bureau. This Bureau is a counseling agency set up by the University as a personnel service open to all students. Testing Bureau cases are given a rather extensive testing program including a series of personality scales. During the summer of 1939, 198 freshman girls came to the Bureau for guidance prior to registration in the University. From this group the experimental group was further selected by the requirement that the research be done on asocial girls as indicated by personality test scores and activity records.

These conditions help to explain why the general problem of the effect of social guidance becomes quite specific, i e., what is the extent of change, if any, in the measured social adjustments and activity records of "under-socialized" University Testing Bureau freshman girls following counseling on social problems and directed participation in extra-curricular and social activities?

The first step in the attack on this problem was the selection of the sample group. The case records of the 198 Testing Bureau cases were read and a record kept of high school scholarship percentile rank, raw score and percentile rank on the American Council *Psychological Examination*, the *Co-operative English Test*, the *Minnesota Inventory of Social Attitudes*—Forms P and B, the *Bell Adjustment Inventory*—Social, and the *Rundquist-Sletto Inferiority Scale*. In addition, each girl's group and individual activities listed on the Individual Record Form of the Testing Bureau were recorded. These two sections are given in the form of a check list on which the subject is asked to indicate those activities "in which you engage frequently." The group activities include team sports, clubs, church organizations, and group parties, while the individual activities are things done alone or with a single other individual, such as sewing, reading, and tennis.

Those girls who had scores in the lower one half of two of the three distributions of Social Preferences, Social Be-

havior (*Minnesota Inventory of Social Attitudes*—Forms P and B), and group activities were selected as the sample group. These three measures were assumed to give an indication of the preferences, behavior, and previous interest in social activities. It should be noted that in order to get a sample of any size the levels had to be fairly high, running up to the median or even higher. This selection from the 198 Testing Bureau cases yielded a sample of 79 freshman girls. This group was then divided into two random samples of 40 and 39, which became the experimental and control groups. The remaining 119 Testing Bureau cases were used for purposes of comparison.

The treatment of these two groups must be emphasized here since this is the crux of the method. The control group of 39 cases was in no way influenced by this study. These girls were handled in the customary manner by the Testing Bureau. Following a preliminary interview and testing, each girl was assigned to one of the five counselors in the Bureau. The usual counseling interview is chiefly concerned with educational and vocational problems. If the social aspects are of importance, however, they may also be considered. No generalizations can be made concerning the social counseling of the control group except that they were exposed to the "normal" counseling program which might include some social guidance.

The treatment of the experimental group, however, went further in giving all of the members of this group an opportunity to participate in social activities. Nine girls in this group as well as 11 in the control group failed to complete the counseling and retesting program. Seven girls in this group were already participating, and the counselors merely discussed their social interests with them. Four were untreated because the counselors felt that academic activities should take all of their time if they were to continue in school. The remaining 20 were interviewed by the counselors with a special emphasis on social adjustment. Each interview took place at the end of the first quarter of the school year. The investigator consulted with the counselors concerning the activities

EXPLORATORY STUDY OF SOCIAL GUIDANCE

which might appeal to each girl, but the interview was very much an individual affair. Following this contact, the investigator attempted to carry out the suggestions of the counselors by personally introducing the girl to those activities in which she expressed an interest. Active participation was facilitated in every way possible. At the same time every precaution was taken to make the social program normal. Introductions were made to various campus organizations which had been informed that the Testing Bureau had appointed a special counselor to act as liaison officer between the Bureau and the organization. The extra-curricular organization heads did not know that this was in any way a research project, and there is every reason to believe that they gave these girls attention similar to that given any girls recommended by a campus agency. The experimental and control groups both had some counseling, but the experimental group had more than the control group.

At the end of the school year both the experimental and the control groups were retested. After three notices all but 20 girls responded; of the 79 originally selected for study, 31 experimental and 28 control subjects were retested. The tests which were given again were the Bell, the Rundquist-Sletto, the Social Preferences, and the Social Behavior. Also included was an Activity Record covering the freshman year. The tests were given under conditions identical with those of the original testing.

It is important that some consideration be given to the original nature of the control and experimental groups. Using the common t test for the significance of the difference between two means,² the sample group (the 59 cases who were retested) does not differ from the rest of the Testing Bureau cases (the Freshman women Bureau cases which were read but not selected for this study) in mean American Council on Education test score or the *Cooperative English Test* score. It does have a significantly higher mean score than frequently

² t =the difference in means divided by the standard error of that difference.

used norm groups. As would be expected, the experimental and control groups are significantly lower in mean score on the Social Behavior and Preference scales when compared with the rest of the Bureau group, but they do not differ in mean Social Behavior score from college Freshman women norm groups as determined from the norms given by the authors of the test. The difference between the mean Social Preference scores for the sample group and the same norm group is just significant (with the sample group lower). The experimental and control groups differ on the Bell and Rundquist-Sletto from the rest of the Bureau group but not from comparable norm groups. These facts lead to the conclusion that although the experimental and control groups are socially "poorly" adjusted when compared with the rest of the Testing Bureau group, they are not clearly different on personality measures from the comparable norm groups. Thus this study was not confined to a group of extreme deviates in personality scores.

Although the control group has a slightly higher median high school percentile rank than the experimental group, the two groups are remarkably similar in original testing on the six objective measures, and all indicated group and individual activities. It is, therefore, safe to assume that any significant differences on retesting may be attributed to differential treatment. It was also found that the 20 cases who did not appear for retesting did not differ significantly on original testing from the rest of the sample group.

The differences on retesting between the experimental and control groups provide the basis for an estimate of the success of social guidance. It would be desirable to obtain some estimate of the amount of guidance and relate this to the amount of participation, although this was not done here. A simple comparison was made of mean gains. These results and those from the Activity Record, using *t* tests where possible, indicate that:

1. There was a significant mean gain made by both the experimental and control groups on retesting on the Rundquist-Sletto Inferiority, the Social Preferences, and the Social Be-

EXPLORATORY STUDY OF SOCIAL GUIDANCE

havior scales. The control group did not gain on the average on retaking the Bell Social scale, while the experimental group did gain.

2. A comparison of the mean gains made on retesting after 9 to 11 months on these measures by the experimental and control groups shows that on all except the Social Preference scale the experimental group gained significantly more (see Table 1).

TABLE 1
MEAN GAINS MADE ON THE FOUR PERSONALITY MEASURES BY THE EXPERIMENTAL AND CONTROL GROUPS ON RETESTING

Measure	Experimental			Control			t	Pt
	N	Mean Gain	S D.	N	Mean Gain	S D		
Social Beh ..	30	4.40	10.12	28	2.21	12.83	3.65	<.01*
Social Pref. ..	30	6.20	13.00	28	6.18	15.47	.03	>.05
Rund-Sletto ..	28	5.17	7.55	26	1.85	6.81	7.64	<.01*
Bell-Social ..	29	3.28	5.95	26	0.00	7.43	8.77	<.01*

*Significant

There is also some indication that the gain is greater for the members of the experimental group who were given the most guidance. The fact that the Social Preference scale shows an insignificant difference in mean gain for the two groups suggests that social guidance has an effect on the actual social behavior or amount of social activity but does not affect social preferences.

3. At the beginning and at the end of the experimental period the counselors rated the members of the experimental group on a rough scale of social adjustment. Only seven per cent of the group was rated lower on second rating and 38 per cent rated higher. There is no comparable measure for the control group, so the significance of this gain is difficult to interpret.

4. The experimental and control groups encircle about the same number of individual activities on retesting, but the experimental group encircles more group activities than the control group on retesting.

5. The experimental group reports more hours per week spent in extra-curricular activities than the control group and more offices and committees in these activities.

6. The experimental group indicates on a rating scale that they want to participate in fewer additional activities, think that they have made more friends, feel that they have participated in more activities compared with high school, and have a better opinion of the extra-curricular and social program on the campus than the control group.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

7. Both the experimental and control groups feel that they are in fewer activities but have made more friends in college than in high school.

All of these findings combine to indicate that, from this small sample, social guidance and directed participation in extra-curricular activities improve the "social adjustment" of freshman girls as measured by personality scales and a questionnaire. Not only do the girls in the experimental group make greater mean gains, but they feel that they have more friends, participate in more activities, and are less critical of the social program than the control group. A treatment that makes people feel better satisfied with their social life is certainly worthy of further consideration. The problem was, however, essentially an investigation of a method and as such the results should be emphasized only as a justification for the further use of the method.

REFERENCES

1. Beaumont, H. "The Evaluation of Academic Counseling", *Journal of Higher Education*, X, (1939), 79-82, 116.
2. Brown, Clara. "A Social Activities Survey", *Journal of Higher Education*, VIII, (1937), 257-265.
3. Burks, F. W. "Some Factors Related to Social Success in College", *Journal of Social Psychology*, IX, (1938), 125-140.
4. Chapin, F. Stuart. *Extra-curricular Activities at the University of Minnesota*. Minneapolis: University of Minnesota Press, (1929).
5. Livingood, F. G. "Directed Extra-curricular Activities and Adjustments", *Mental Hygiene*, XX, (1936), 614-623.
6. Mallay, H. "A Study of Some of the Factors Underlying the Establishment of Successful Social Contacts at the College Student Level", *Journal of Social Psychology*, VII, (1936), 205-228.
7. Tuttle, H. S. "The Campus and Social Ideals", *Journal of Educational Research*, XXX, (1936), 177-182.
8. Williamson, E. G. and Bordin, E. S. "Evaluating Counseling by Means of a Control-group Experiment", *School and Society*, LII, (1940), 434-440.
9. Williamson, E. G. and Darley, J. G. "The Measurement of Social Attitudes of College Students. II. Validation of Two Attitude Tests", *Journal of Social Psychology*, VIII, (1937), 231-242.
10. Wrenn, C. G. *The Evaluation of Guidance*, Purdue University: Studies in Higher Education, No. 37, (1940), 51-61.

NEW TESTS*

Cooperative Chemistry Test for College Students, by B. Clifford Hendricks, B. H. Handorf, O. M. Smith, Chris P. Keim, Rufus D. Reed, Alexander Calandra, Ralph W. Tyler, and Fred P. Frutchey. Form 1942. Part I, Information and Vocabulary; Part II, Problems and Equations; and Part III, Scientific Method. Time, 90 minutes. 10 to 99 copies 6½c; 100 or more copies 6c; specimen set 25c. Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York City.

Cooperative English Test, by Geraldine Spaulding and Frederick B. Davis. 1942. Form S. Test A, Mechanics of Expression; Test B1, Effectiveness of Expression (Lower Level); Test B2, Effectiveness of Expression (Higher Level); Test C1, Reading Comprehension (Lower Level); Test C2, Reading Comprehension (Higher Level). Time, 40 minutes for each test. 10 to 99 copies 5½c; 100 or more copies 5c; specimen set 25c. Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York City.

Cooperative French Test, by Geraldine Spaulding, Laura Towne, and Sarah Wolfson Lorge. 1942. Form S, Lower Level for use in the first two years of high school or the first year of college, Higher Level for use with students who have had more than two years study of French in high school or more than one year in college. Part I, Comprehension; Part II, Grammar; and Part III, Civilization. Time, 80 minutes. 10 to 99 copies 6½c; 100 or more copies 6c; specimen set 25c. Part I available as separate booklet, 10 to 99 copies 5½c; 100 or more copies 5c, specimen set 25c. Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York City.

Cooperative Italian Test, by Peter Riccio and Anthony Cuffari. 1942. For students who have had two semesters or more of study of Italian. Experimental Form S. Time, 70 minutes. Part I, Reading; Part II, Vocabulary; and Part III, Grammar. 10 to 99 copies 6½c; 100 or more copies 6c; specimen set 25c. Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York City.

Cooperative Latin Test, by Harold V. King and Geraldine Spaulding. 1942. Form S, Lower Level to cover beginning Latin and Caesar;

*Prepared by Jane Gilbert.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Higher Level for use with students who have completed one semester or more of study beyond Caesar. Part I, Comprehension; Part II, Grammar; and Part III, Civilization. Time, 80 minutes. 10 to 99 copies 6½c; 100 or more copies 6c; specimen set 25c. Part I available as separate booklet, 10 to 99 copies 5½c; 100 or more copies 5c; specimen set 25c. Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York City.

Cooperative Test in Secondary School Mathematics (Higher Level), by Margaret Martin, William Mollenkopf, Radcliffe W. Bristol, William S. Litterick, and Carroll G. Ross. 1942. Form S. For grades 10 to 12. Time, 80 minutes. 10 to 99 copies 6½c; 100 or more copies 6c; specimen set 25c. Published by the Cooperative Test Service, 15 Amsterdam Avenue, New York City.

Interest Inventory for Elementary Grades, by Mitchell Drees and Elizabeth Mooney. 1941. Time, about 30 minutes. 5c each; manual 15c; specimen set 25c. Published by the Center for Psychological Service, George Washington University, Washington, D. C.

Meier Art Judgment Test, by Norman Charles Meier. Revised 1942. Grades 7 through adult. Time, about 45 minutes. Test books 75c; \$3.50 for 5; \$6.25 for 10c; 55c in lots of 25; record sheets 2½c; 2c per 100; manual 10c; sample set 90c. Published by the Bureau of Educational Research, State University of Iowa, Iowa City, Iowa.

Otis Classification Test, by Arthur S. Otis. Revised 1941. Forms R, S. and T. For grades 4 to 8. Time, 30 minutes for each part. Hand- and machine-scored. \$1.25 per 25; specimen set 30c. Published by the World Book Company, Yonkers-on-Hudson, New York

Pintner-Durost Elementary Test, by Rudolf Pintner and Walter N. Durost. For grades 2, 3, and 4. Form A, Scale 1 (Picture Content) and Scale 2 (Reading Content). \$1.35 per 25 for Scale 1; \$1.20 per 25 for Scale 2; specimen set (Scale 1 and Scale 2) 30c. Published by the World Book Company, Yonkers-on-Hudson, New York.

Preference Record, by G. Frederic Kudei. 1942. Form BB for self-scoring; Form BM for machine-scoring. For high-school and college students and adults. Time, about forty minutes. Test booklets 25c; answer pads 5c; profile sheets \$1.25 per 100; specimen set 25c.

NEW TESTS

Published by Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois.

Purdue Placement Test in English, by J. H. McKee, G. S. Wykoff, and H. H. Remmers. 1941. For high school seniors and college freshmen. Time, about 35 minutes. Form C, \$1.65 per 25; separate answer sheets 75c per 25. Published by Houghton Mifflin Company, 2 Park Street, Boston, Massachusetts.

Terman-McNemar Test of Mental Ability, by Lewis M. Terman and Quinn McNemar. 1942. For grades 7 to 12. Time, 40 minutes. Forms C and D. \$1.25 per 25; specimen set 20c. Published by the World Book Company, Yonkers-on-Hudson, New York.

Study-Habits Inventory, by C. Gilbert Wienn. Revised 1941. For grade 12 and college. \$1.25 per 25, \$3.50 per 100; \$2.50 per 100 for 1000 or more. Published by Stanford University Press, Stanford University, California.

Test of Practical Judgment, by Alfred J. Cardall. 1942. For 12th grade level and above. Time, about 45 minutes. Hand- or machine-scored. 10c each; specimen set 25c. Published by Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois.

The World Test, by Charlotte Buehler and Gayle Kelley. 1941. To measure emotional problems. For clinical use with children 5 to 11. Time, about 20 minutes. Complete test materials, manual, and 25 record forms, \$60.00. Published by the Psychological Corporation, 522 Fifth Avenue, New York City.

MEASUREMENT ABSTRACTS*

Bellows, R. M. "Procedures for Evaluating Vocational Criteria." *Journal of Applied Psychology*, XXV (1941), 499-513.

The fact that the basic vocational criteria used in the evaluation of predictive instruments are fallible is generally neglected. The source of fallibility may lie in such factors as (1) illicit use of predictive information giving previous knowledge of psychological test scores or other performance ratings; (2) artificial limitations of production brought about by physical conditions influencing output of work; (3) differential experience or training. To overcome the influence of criterion contamination several checks are recommended and evaluated. Knowledge of the future validity of a predictor is impossible because of various changes in the situation. No single procedure for criterion evaluation is adequate, which suggests that indices of validity are largely determined by the degree of fallibility of the criterion, and that the interpretation of such indices is dependent upon knowledge of the criterion used in validation. *L. Bouthilet.*

Buros, Oscar K. (editor) *The Second Yearbook of Research and Statistical Methodology*. Highland Park, New Jersey, Gryphon Press. 1941.

This yearbook has been compiled in an effort (a) to make students and teachers of statistics aware of inaccuracies and the inadequacy of much current statistical literature and information, (b) to serve as a source for selection of textbooks with discrimination, (c) to evaluate weak and strong points of statistical books, (d) to point out current developments in monograph and textbook writing and criticism, (e) to acquaint statistical workers with the broad applications of statistical work in many fields, (f) to present different points of view among students of statistical theory, (g) to improve the quality of such book reviews by more careful choice of reviewers and by stimulating reviewers not to review books which they cannot appraise adequately.

The editor has greatly increased the scope of this volume over an earlier one, including 1,652 review excerpts from 283 journals. An attempt has also been made to list books on research methodology in specific fields, although this list is by no means inclusive. However, this yearbook represents a significant contribution to the field of methodology and should make workers in this field more acutely aware of current developments. *Jane Gilbert.*

Cronbach, L. J. "An Experimental Comparison of the Multiple True-False and Multiple Multiple-Choice Tests." *Journal of Educational Psychology*, XXXII (1941), 533-543.

*Edited by Forrest A. Kingsbury.

MEASUREMENT ABSTRACTS

Two subject-matter tests, one in multiple true-false form, and the other in multiple multiple-choice form were administered to 57 and 60 students, respectively. The former consists of multiple-choice items in which each alternative is marked true or false by the student; the latter, of similar items in which only correct alternatives are marked. Results showed the two forms were essentially equivalent. The hypothesis is advanced that the tendency to mark uncertain items "true" may be a personality trait which may influence the validity of true-false test scores. *L. Brdsall.*

Ewart, E., Seashore, S. E., and Tiffin, J. "A Factor Analysis of an Industrial Merit Rating Scale." *Journal of Applied Psychology*, XXV (1941), 481-486.

In order to determine how many traits actually influence the ratings, tetrachoric intercorrelations were computed for ratings on a twelve-trait scale constructed for use in a large industrial plant. This correlation matrix was factored by Thurstone's centroid method, and the factors rotated for simple structure. Three factors were obtained: I, a general factor, termed "ability to do the present job," accounts for most of the total variance of the scale. Factor II represents knowledge or skill over and above the requirements for the specific job. Factor III is on the variable "health." Factors I and III are orthogonal while Factors I and II are oblique. *K. S. Yum.*

Foulano, G. and Pintner, R. "Selection of Upper and Lower Groups for Item Validation." *Journal of Educational Psychology*, XXXII (1941), 544-549

Two sets of data from the *Study Habits Inventory* and the *Home-Background Survey Test* have been subjected to item validation, using five different methods of selecting upper and lower groups. The authors conclude that for a simple and rapid, rough-and-ready method of validation of test items of the inventory type, the upper versus lower 27 per cent method is preferable, even though distributions are more or less non-normal. The other upper versus lower methods studied were 50 per cent, 33⅓ per cent, 16 per cent, and 7 per cent. *K. S. Yum.*

Gilman, W. A. and Gray, D. E. "Guessing on True-False Tests." *Educational Research Bulletin*, XXI (1942), 9-12.

The attempt to penalize guessing by subtracting the number of wrong answers from the number of correct answers is ineffective. This is clear from the study of a case in which there are n pure guesses. Theoretically the student would have $\frac{n}{2}$ correct answers and $\frac{n}{2}$ incorrect answers; hence the increment of $\frac{n}{2} - \frac{n}{2}$ would leave his grade on other

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

items unaltered. In practice, pure guessing rarely exists. The possession of partial knowledge gives the student better than a fifty per cent chance; hence it is to his advantage to guess on tests thus scored. *George W. Boguslavsky.*

Gowdon, C. H. "The Revised Stanford-Binet Scale Applied as a Point-Scale." *Journal of Applied Psychology*, XXV (1941), 660-671.

Form L of the Revised Stanford (from Year VI up) has been arranged as a point-scale. By testing each subject only with that range of tests limited by 5 consecutive successes below the first failure and 5 consecutive failures above the last success, a very reliable mental age is obtained. Rescoring the records of 440 children of the usual clinical types yields I. Q.'s which correlate .976 with regular Stanford-Binet I. Q.'s. For eleven mental age levels the average saving in number of tests given was about 35 per cent, with I. Q. variations not to exceed 5 points in 9 of every 10 cases. *F. A. Kingsbury.*

Harding, J. "A Scale for Measuring Civilian Morale" *Journal of Psychology*, XII (1941), 101-110.

Out of a list of 59 items given to two criterion groups, high morale and low morale, 20 items were chosen to form the present morale scale. Each item is included in one of the four clusters: a. an attitude of confidence in the broad framework of capitalist democracy; its opposite, cynicism; b. an attitude of tolerance for various groups, c. an attitude of realism as opposed to wishful thinking; d. an attitude of assertive idealism in international affairs. Scoring for each item is on a five-point scale. Thus "total morale scores" may be computed. *Louise Grossnickle.*

Heston, J. C., and Cannell, C. F. "A Note on the Relation Between Age and Performance of Adult Subjects on Four Familiar Psychometric Tests." *Journal of Applied Psychology*, XXV (1941), 415-419.

Vocabulary Tests from Form L of the Revised Stanford-Binet Scale, Knox Cubes, Porteus Mazes and Ferguson Form Boards Tests were given to members of borrower families of the F.S.A. in Ohio, Maine, and Missouri. The data include 643 cases, 375 men and 268 women, all white. The age range for men was 15 to 76, and for women, 15 to 72, with medians at 37.5 and 35.0 years respectively. Two contrasting tendencies are noted on the age curves of scores of these tests. On the vocabulary test there is a rapid increase from age 15 to 20, then a slight rise up to 55, where a small drop occurs; while on the performance tests a rapid decline seems to be a characteristic tendency. *K. S. Yum.*

Jones, H. E. "Seasonal Variations in I. Q." *Journal of Experimental Education*, X (1941), 91-99.

MEASUREMENT ABSTRACTS

A study of 19 comparisons of fall-to-spring versus spring-to-fall I. Q. changes in children of preschool age revealed that 18 of the 19 comparisons show a greater gain over the winter interval than over the summer interval. Four alternative hypotheses were considered:

1. Seasonal variations in the testers.
2. Seasonal variations in test performances.
3. The dependence of performance on seasonal variations in the child's activity.
4. The effect of seasonal variations on mental and physical growth.

George W. Boguslavsky.

Katz, Evelyn. "The Constancy of the Stanford-Binet I. Q. From Three to Five Years." *Journal of Psychology*, XII (1941), 159-182.

The Brush Foundation of Western Reserve has the records of 308 children of high socio-economic level, tested at six-month intervals from three to five years of age. "Test-retest correlations range from .533 to .765, the size of the correlations being unrelated to age but inversely related to the interval between tests." "The group as a whole shows a small increase in I. Q. with age." Large gains and losses of 20 or more points are more frequent over the longer intervals of time and for the younger ages. They are present in approximately 10 per cent of the test-retest comparisons, and occur for 40 per cent of the children. "These frequent fluctuations should probably be regarded as typical of children between three to five years who come from families of superior socio-economic status." *Helen M. Wolfe.*

Lindquist, E. F. *A First Course in Statistics*. Boston, Houghton Mifflin Company. 1941. 240pp.

This elementary statistics textbook presents a well-organized approach to the problem of measurement. An accompanying workbook has been designed to help the student integrate the theoretical approach with actual practice in applying these principles. The topics presented are as follows: frequency distribution, percentiles, graphical representation of frequency distributions, measures of central tendency, measures of variability, the nature of the normal curve, sampling error theory, standard measures and methods of combining test scores, correlation theory, and correlation techniques applied in the evaluation of test materials. *Jane Gilbert.*

Morrow, Robert S. "An Experimental Analysis of the Theory of Independent Abilities." *Journal of Educational Psychology*, XXXII (1941), 495-511.

"Eighty relatively homogeneous male college students were given

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

in a random manner" 23 subtests of standard tests of intelligence, artistic judgment, and clerical, mechanical, and manipulative ability. The correlations were analyzed by the "center of gravity" method into four factors. The factors were not isolated and were difficult to interpret. "By virtue of these findings," states the author, "it would appear that the Spearman and Thurstone theories are inadequate for explaining the relationships expressed in this study. Rather, one must conclude with the hypothesis that the abilities here tested are not disparate and static abilities, but that they are, instead, functional and dynamic relationships within the total personality." *Helen M. Wolfe.*

Reed, H. B. "The Place of the Bernreuter Personality, Stenquist Mechanical Aptitude, and Thurstone Vocational Interest Test in College Entrance Tests." *Journal of Applied Psychology*, XXV (1941), 528-534.

This investigation proposed to investigate the inter-relationships between the Bernreuter, Stenquist, and Thurstone tests and scholastic achievement in order to find the place of such tests in a battery of college entrance examinations. The Bernreuter scores were also compared with teachers' ratings of traits of the same name as those in the test. Results show that there was little or no relationship between the three tests and scholastic achievement. It is concluded that the tests are of little value for guidance in choice of college courses, although the usefulness of the tests for other purposes was not investigated. *L. Bouthilet.*

Robinson, Frances P. *Diagnostic and Remedial Techniques for Effective Study*. New York, Harper and Brothers. 1941. 318pp.

This handbook has been evolved as a result of the author's experience at the State University of Iowa and an extensive how-to-study program at Ohio State University. The major emphasis in this book has been placed on diagnostic tests which are based on research analyses of college work and student errors rather than standard academic organization. The types of areas measured include study habits, reading skill, skill in use of academic resources, knowledge of fundamental processes and background knowledge, health, vocational planning, social adjustment, personal problems and motivation. All materials necessary for test administration and scoring are included in this book. Comparable retests are also available to help the student evaluate his improvement and to see the nature of his remaining problems. The book cannot be used independently by a student, but it should form a working basis for individual counseling and delineation of specific areas in which remedial treatment is indicated. *Jane Gilbert.*

Shuttleworth, F. K. "Sampling Errors Involved in Incomplete Returns to Mail Questionnaires." *Journal of Applied Psychology*, XXV (1941), 588-591.

MEASUREMENT ABSTRACTS

There has been little attempt to determine the sampling errors due to incomplete returns of mail questionnaires. The only adequate check is to compare incomplete returns with complete returns. In a study of the employment status of certain university alumni, it was found that serious sampling errors were involved, the earliest returns coming from the more successful alumni. The conclusion is drawn that each questionnaire situation needs intensive study, which should include a complete return from at least a portion of the total population. *L. Bouthilet.*

Sloan, W. and Sharp, A. A. "A Note on Interpolation of Kent Oral Emergency Test Scores into Mental Age Years and Months." *Journal of Applied Psychology*, XXV (1941), 592-594.

The method consists of dividing equally the 12 mental age months in each year so that the first point at each year level falls exactly on that year. A corresponding column of I. Q.'s for adults is given with the chronological age of sixteen as a constant divisor. *K. S. Yum.*

Stalnaker, J. M. "A Note on the Computation of Y Values for Integral Values of X, when Y is a Linear Function of X." *Journal of Educational Psychology*, XXXII (1941), 559-560.

The author reports a method for rapid and accurate determination of converted scores for a large number of raw scores, with the aid of accounting machines and punched-card methods. The method demands a minimum of hand labor. The procedure is applicable to any situation where one set of scores is to be transmuted into any set of scores, providing the two sets are in a linear relationship, and the one variable changes in unit steps. *K. S. Yum.*

Super, D. E., and Roper, S. A. "An Objective Technique for Testing Vocational Interests" *Journal of Applied Psychology*, XXV (1941), 487-498.

A technique developed for testing vocational interests objectively is described. Pictures and films depicting different phases of various occupations are used. The assumption is made that memory for what is seen will be greatest in the field of greatest interest. Methods of validation are described. The test of interest in nursing was administered to 35 nurses and 111 high-school students, 36 of whom planned to enter nursing. Intelligence, previous knowledge, and success in nursing school influence the scores slightly or not at all. The present test showed no correlation with the *Strong Vocational Interest Blank*. The authors conclude the two are equally valid, but that the former measures degree of interest, whereas the latter compares the interests of subjects and those in the field. *L. Birdsall.*

Triaxler, A. E. "The Reliability of the Bell Inventories and Their Correlation with Teacher Judgment" *Journal of Applied Psychology*, XXV (1941), 672-678.

Scores of 43 high-school pupils on the Bell *Adjustment Inventory* and the Bell *School Inventory* have been correlated by the split-half method. All the reliability coefficients are above .80, and some of them are close to or above .90. Correlations between the scores and the ratings by teachers and counselors on 33 pupils have been obtained. Four of the six correlations are statistically significant. However, the correlations, being low, fail to substantiate the validity of the inventories. The author suggests that we should have a criterion that will be much more defensible than a rating scale. *Louise Grossnickle*.

Yum, K. S. "Primary Mental Abilities and Scholastic Achievements in the Divisional Studies at the University of Chicago." *Journal of Applied Psychology*, XXV (1941), 712-720.

What particular combination of primary mental abilities is required for success in the divisional studies of the physical, biological and social sciences? The scores of 110 University of Chicago juniors were examined. "According to the critical ratios, there apparently exists no significant difference between the biological and social science groups." The mean profile of the physical science group (but not the total score) is significantly different from the other two. Induction distinguishes physical science men from biological science men, and deduction, space, and induction distinguish physical science men from social science men. The correlations of the factors with grades range from $-.17$ to $+.52$. "In general, the verbal, inductive reasoning, and deductive reasoning factors seem to correlate better with scholarship." *Helen M. Wolfe*.

MEASUREMENT NEWS

The Personnel Procedures Section, formerly the Personnel Research Section, of the War Department has developed in recent months a variety of classification, special aptitude, and achievement tests for the use of the Army. The section is currently interested in the selection of officer candidates and military specialists and in the training of physically and mentally limited men.

Among the officers on active duty with the section are Major Morton A. Seidenfeld, formerly of the National Tuberculosis Association; Lt. Donald E. Baier, on leave from the Mental Hygiene Bureau of the New Jersey State Hospital; and Lt. T. W. Harrell, who was in charge of research for the section in a civilian status. Captain Sidney Adams, formerly of the Employment Section of the Tennessee Valley Authority, has left the section for duty in the field. Among the civilian personnel of the section are: Dr. Clyde H. Coombs, formerly of the University of Chicago; Dr. Louise R. Witmer, on leave from Florida State College for Women; Dr. Bronson Price, formerly of Ohio State University; Dr. Reign H. Bittner, also formerly of Ohio State University; Mrs. Ruth D. Churchill, formerly of the University of Minnesota; Dr. Alvin C. Euich, formerly of Stanford University; and Mr. Howard Uphoff, formerly of the U. S. Civil Service Commission.

Schools interested in building pupil morale for meeting war hardships will be interested in a "Test on the Effects of War" designed for the study of pupil morale and to identify war problems about which further instruction is needed. The test, prepared by Dr. Lee J. Cronbach, has been released by the School of Education of the State College of Washington, Pullman, Washington. The test is planned for grades 10, 11, and 12, but may be used at higher levels. Seventy statements about conceivable future developments are presented, and the pupil is required to respond by indicating how likely he thinks each effect is. Responses are analyzed to determine how optimistic or pessimistic each pupil is. Since good morale depends on a realistic outlook and planning for future developments, both the highly optimistic or complacent pupil, and the highly pessimistic, panicky pupil, are pointed out as cases for individual guidance. An item analysis of the responses of the group indicates those particular war problems about which pupils appear poorly informed.

The test is being made available as a professional service on a non-commercial basis to interested schools. For greatest value in planning the school program during wartime, the test should be given as early as possible. Question sheets, which may be used any number of times, sell for one cent apiece. Answer sheets, one of which is needed for

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

every pupil tested, sell for five cents each. This charge covers the cost of producing the test and of a complete scoring service. All papers are scored, analyzed, and interpreted by the State College without additional charge.

The test has been standardized on nearly two thousand pupils in the State of Washington, tested during January and February, 1942. The reliability of the Optimism score, on which the principal interpretations are based, is .77.

The Committee of Examinations and Tests, Division of Chemical Education, of the American Chemical Society, has announced that the 1942 Cooperative Chemistry Test will be available by April first. Inquiries should be addressed to the Cooperative Test Service, 15 Amsterdam Avenue, New York City.

The accumulation of data and experience in recent years has had the effect of modifying the concept of what the test should measure. As a result of extensive discussion at a conference held at the University of Chicago last June, the 1942 Form of the test is considerably different from the tests of the past four years. The test has been administered in a preliminary form to determine the difficulty and validity of each item. A brief description of the test follows:

Part I. *General Knowledge and Information.*

This section is based on knowledge of or *acquaintance with* important facts, definitions, laws and theories of chemistry. Historical events and application of chemistry to the social and economic world are represented.

Part II. *Application of Principles.*

This part attempts to measure the ability to solve numerical problems, to balance equations, and to make quantitative predictions by the application of chemical principles.

Part III. *Scientific Method.*

This section is concerned with the understanding of the relation of observation, definitions, laws, theories in the scientific procedure. The relation of theory to experiment is represented, as well as the ability to interpret chemical data.

Part IV. *Knowledge of Laboratory Technique and Procedure.*

This new section is included in the effort to measure acquaintance with the laboratory and knowledge of "correct" procedures. It does not attempt to measure skill or technique *per se*.

MEASUREMENT NEWS

The committee which is sponsoring this test is comprised of the following members of the Division of Chemical Education:

B. Clifford Hendricks, University of Nebraska.

Rufus D. Read, New Jersey State Teachers College.

Ed. F. Degering, Purdue University.

Laurence S. Foster, Brown University.

Earl W. Phelan, Georgia State Womans College.

Theodore A. Ashford, University of Chicago.

Otto M. Smith, Oklahoma A and M College, *Chairman*.

The Annual Report of the Scottish Council for Research in Education states that a mass of records — 2,500 of scale L and 350 of scale M — have been collected with a view to standardizing the Terman-Merrill Revision of the Stanford Binet Scale for use in Scotland. It is hoped shortly to produce some evidence as to the suitability of this Revision for Scottish children.

A report on the follow-up of the random sample of 1,000 children and of the high scorers in certain counties who were given the Binet test in the 1932 Mental Survey is awaiting publication. It is interesting to note that an independent analysis of occupations has been made and correlated with each I. Q. group. The relation of occupation to age and to class on leaving school has been worked out with respect to both initial and final occupations, that is, to those occupations entered upon leaving school and to those held for not less than one year immediately before the close of the survey. A geographical analysis, based on the Four Cities, urban areas excluding the Four Cities, and rural areas, has also been made. The Annual Report states that so far as can at present be ascertained the correlation between intelligence and initial occupation does not appear to be very high, but this relation is closer by the time the occupation held at the close of the follow-up is entered.

Another report awaiting publication covers the results of an inquiry into methods of forecasting, at the qualifying stage, the pupil's later success. The methods considered are the traditional examination, scholastic tests, an intelligence test and teacher's estimate. It appears that the best combination, productive of the least number of misfits, is an intelligence test, an examination, and teacher's estimate "scaled."

The Psychological Classification and Research Sections of the Army Air Forces have established three Psychological Research Units. These units are located at Maxwell Field, Alabama, Kelly Field, Texas, and Santa Ana, California, and are headed respectively by Laurence F. Shaffer, Robert T. Rock, Jr., and J. P. Guilford.

All aviation cadets are given psycho-motor and group tests for the purpose of classifying them for various duties in the aircrew. In addi-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

tion to administering these tests, the units do some research on the general problem of determining the aptitudes needed for different aircrew duties, as well as develop methods for the prediction of success in such duties.

These units are staffed by a group of officers and enlisted men. All the officers are well-qualified psychologists. Most of the enlisted men have done some graduate work in psychology and in addition have had some experience either in using psychological laboratory equipment or in the development, use, and validation of psychological tests. Periodically some qualified enlisted men are recommended for officer candidate schools. Successful completion of such schools leads to a commission.

Men interested in enlisting for such positions should send the following information to the Army Air Forces, Office of the Air Surgeon, War Department, Washington, D. C.: (1) full name, (2) date and place of birth, (3) local board number and order number, (4) four personal references, and (5) complete work and educational histories, including a detailed description of specialized training in psychology. Individuals who expect to be inducted into the service soon and who desire to be considered for assignment to work in psychology, should, in addition to the previous information, include (6) probable date of induction, stating whether notification of date of induction has been received and (7) probable place of induction.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Volume II

JULY, 1942

Number 3

THE EXAMINERS OFFICE OF THE UNIVERSITY SYSTEM OF GEORGIA	233
<i>F. S. Beers</i>	
LEVELS OF COMPETENCE IN COUNSELING—A POST-WAR PROBLEM FOR STUDENT PERSONNEL WORK IN SECONDARY SCHOOLS.	243
<i>Milton E. Hahn</i>	
A STUDY OF SOME LOCAL FACTORS AFFECTING STUDENTS' SCORES ON THE MINNESOTA PERSONALITY SCALE.	257
<i>Betty M. Horne and W. C. McCall</i>	
THE PLACE OF APTITUDE TESTING IN THE PUBLIC SCHOOLS	267
<i>Donald E. Super</i>	
EFFECT OF ENGINEER SCHOOL TRAINING ON THE SURFACE DEVELOPMENT TEST	279
<i>Ruth D. Churchill, Jeanne M. Curtis, Clyde H. Coombs, and Thomas W. Harrell</i>	
AN AID TO STUDENT COUNSELORS.	281
<i>Ralph F. Berdie</i>	
A COMPARISON OF THE HUMAN BEHAVIOR INVENTORY WITH TWO OTHER PERSONALITY INVENTORIES.	291
<i>Abraham Sperling</i>	
INTRA-INDIVIDUAL DIFFERENCES VERSUS INTER-INDIVIDUAL DIFFERENCES IN MOTOR SKILLS	299
<i>William A. Owens, Jr</i>	
NEW TESTS	315
MEASUREMENT ABSTRACTS	317

Copyright, 1942, by
SCIENCE RESEARCH ASSOCIATES

PRINTED IN THE UNITED STATES OF AMERICA

THE EXAMINERS OFFICE OF THE UNIVERSITY SYSTEM OF GEORGIA

F. S. BEERS
Social Security Board

THE UNIVERSITY SYSTEM OF GEORGIA is unique among the states. It is a centrally administered, governmentally supported organization of 15 colleges now completing its first decade. Whether state-supported higher education so conceived and so administered can and should endure is a question which is fittingly being tried out, as it were, in "the oldest chartered state university" and its branches.

Before 1931 there were 25 state-supported colleges in Georgia, with a grand total of 365 college trustees. Each college operated as a unit, appealed to the legislature for financial support in competition with the other colleges, arranged its curriculum and its administration as it saw fit, and ordered its affairs to please itself. The older and stronger of the colleges used as their chief defensive weapon a policy of paring down or reducing in value the credits earned at the younger and weaker colleges, thus discouraging enrollment at these institutions and exacting tribute of students who transferred from them.

In one stroke the Reorganization Act of 1931 swept this scramble into the discard. Ten colleges were abolished,¹ a single Board of Regents replaced the 365 local trustees, and a chancellor was set up as chief administrative officer. In the Chancellor and the Board of Regents was vested the authority

¹The colleges surviving the reorganization were. The University of Georgia with its School of Medicine, The Georgia School of Technology, two senior colleges for women, one college for teachers, seven junior colleges, and three colleges for Negroes. The average annual enrollment in regular session is about 12,000 students.

for setting the educational and financial policies of the system of colleges and for reviewing the activities of colleges individually, as they might add to or detract from the effectiveness of service to the state.

As part of the reorganization it was recommended that "at an early date there should be added to the Chancellor's office an . . . officer properly trained in educational and statistical techniques [who] should be charged, under the supervision of the Chancellor, with the necessary duty of assembling, analyzing, and interpreting the regular and special reports of the operations of the several branches so as to make continually available in proper form for the Board of Regents that general information and other specific data upon which the Board may base its actions."

An office for this purpose was established in 1934 by order of the Regents and was located at the University,² that being considered the hub of the academic wheel whose circumference is the state system of higher education. The Regents wisely provided this new office with the nucleus of a bureau of standards against which educational accomplishments and experiments could be measured and from which administrative policies of individual colleges could be, directly or indirectly, evaluated.

This provision included a basic curricular pattern of ten courses representative of general education, which was required in all the colleges, the content of the courses having been determined by the faculties of the colleges in a series of conferences.

To provide for the effective administration of these courses, information for their frequent revision, and a guarantee that equal achievement on the part of students regardless of college should be given equal credit, with right of transfer of credits without let or hindrance, the Regents authorized state-wide examinations on these courses and common interpretation of scores made by students taking them.

²The University is in Athens; the Chancellor's office is in the State Capitol, Atlanta.

EXAMINERS OFFICE OF GEORGIA SYSTEM

Administrative responsibility for this policy was assigned by the Board to the new office it had set up, which later, by general acceptance, came to be known as the "Examiners Office." Supervising and administering course examinations, however, were intended to be no more than partial bases of operations for more important duties and obligations of the office.

The primary bureau of standards, consisting of 10 survey courses generally required, was augmented by authority to make use of a variety of devices for gauging the effectiveness of the program, among them measures of the relative quality of students electing to attend and those not electing to attend college, together with ways of improving selection; analysis of the physical well-being of students and its relation to mental acuity; evaluation of the amount of cultural background, skill, and intellectual power that the college environment provides, with applications to the individual problems of students through educational and vocational guidance.

From the general framework and techniques of analysis that are employed in the attack upon these problems have come numerous adaptations that provide partial, and often rather full, answers to such questions as the relative cost to the state of general and special education, the relative effectiveness of each as judged by administrators, as observed in student opinion and experience, and as measured against outside criteria; optimum size of class enrollments; whether education conceived and practiced as a purely personal matter between instructor and student tends to crystalize and crumble more or less rapidly than when it is broadly administered and variously supervised as, for example, under central as against local administration, or under divisional as against departmental jurisdiction. Nearly all of these problems revolve on the one hand around individual college prerogatives, and on the other around obligations of *the colleges*, collectively, to the state.

How successful has this venture toward a university system proved itself to be?

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

It is perhaps too early, after the lapse of approximately a decade only, to judge fairly whether the experimental attempt to establish the University System and maintain it through a research program will ultimately make a positive contribution to policy in higher education generally, as the founding of the first State University did so well over a century and a half ago.

Of the successes or failures, those in administration are, of course, the most difficult to appraise. Any issues that point up a central administration as superseding local or college jurisdictions are bound to generate an appreciable amount of heat and to invite emphatic assertion of "states' rights." Hence, it is to be expected that patterns of incandescent light will flicker frequently, as they have, among the colleges of Georgia and will wax in strength within the faculties of individual colleges and departments as well. Attention must thus inevitably be divided, often on very plausible and sincerely held grounds, between these things and many admirable accomplishments of central administration such as are annually cited by the Chancellor in his report to the Governor.

But in the less controversial sphere of service and research, it can be claimed with confidence that many of the techniques employed in the experiment toward a university system have not only served their immediate purpose well but also have proved useful in helping to set administrative policies.

Course examinations in the basic curriculum have been the key to assembling data on the effectiveness of the educational program. These courses are set up on a five-hour, quarterly plan. There are three in the social studies, two each in the biological and physical sciences, two in humanities, and one in mathematics. Two in elementary English were added to this group for examination purposes in 1936; and two in chemistry in 1940. Since the war emergency, additional courses in mathematics and physics are also being included for the men students.

Quarterly from 500 to 2,500 person-examinations are administered in each of the basic courses, with an average per

EXAMINERS OFFICE OF GEORGIA SYSTEM

course approximating 1,000. Over the period of a year about 200 teachers take part in the instructional and examining programs.

Administrative procedures for these programs are designed to furnish a framework within which the individual talents of teachers not only may be protected but also may be given direction. These procedures may be summarized as follows:

1. Conferences and committees to formulate the aims of courses and the content-outline of examinations, and to consider the limitations imposed upon the objectives of examining by the average student to be served and by the extent of variability from this average of other students who likewise are to be judged on examination results;
2. Participation by teachers in these general formulations, in the types of questions or tasks that will be required of persons to be examined, and in the final selection of materials to be included in an examination;
3. Definite and regular assignments in question making for inclusion in examinations, with complete encouragement to offer innovations with respect to both form and content; and
4. An office for analysis, research, and collation of the examining function with respect to construction, administration, and interpretation.

Individual and group conferences are used extensively for the purpose of aiding members of the teaching staff in the preparation and improvement of examination questions. Item analyses are placed at the disposal of committees and instructors, and reports, digests, mimeographed material, and the like are made available for the information of staff members.³

Scaling of examination results is done by the Examiners Office, with the advice of Divisional Heads and after periodic canvassing of faculty opinion. Final grades in course work are assigned by individual teachers by means of a type of

³For example, F S Beers and others, *Some Principles of Examining, with Aids for Consulting Examiners* (University of Georgia Press, 1942), 45 pages.

transmutation to letter grades of the average ranks on class work assembled by teachers and on scores on the final examination.⁴ Final grades are comparable from college to college and among the basic courses. This feature is imperative if the transfer of credits and students is to be effected on a legitimate basis; and in a university system it should take precedence over the more common practice in colleges of subordinating the examining function to the much vaunted "objectives of instruction." It is apparent, however, that the two points of view need not be mutually exclusive. Those who make such a claim tend, perhaps, to think more with their bile than with their brains.

Machine scoring of final examinations is done by the staff of the Examiners Office. From 9,000 to 13,000 answer sheets are scored quarterly in a period not exceeding four days. The method used is unique. Each college alphabetizes its answer sheets immediately after an examination, prepares an alphabetical list of student names in duplicate, packages both, and expresses or brings the packages to the central office. Here the procedure is as follows:

1. The duplicate list of names is inserted in a typewriter set up adjacent to the scoring machine, and a typist is put in charge to record scores;
2. A tally clerk equipped with printed forms is located in front of the scoring machine facing the operator;
3. The scoring machine operator calls each name and score (part or whole as the case may be) but does not record it on the answer sheet;
4. The typist and tally clerk record the scores on their respective forms from the "call" of the machine operator;
5. A calculating machine operator summates the tallies by sections and colleges and runs the scale on the total distribution for the State;

⁴See F. S. Beers and H. M. Cox, "Measurement or Marking?" *Journal of the American Association of Collegiate Registrars* (April, 1938).

EXAMINERS OFFICE OF GEORGIA SYSTEM

6. Names, raw scores, and the scale for transmuting scores into grades are put in an envelope and mailed special delivery to each college dean.

On the average, the total possible score per examination is 150 points, although some tests may have a possible total of more than double this figure. Rescoring has shown small errors, ± 1 point, to be characteristic of about 5 per cent of answer sheets. Only very occasionally do large errors occur. As a check on these, deans are instructed to call for rescoring whenever a student's displacement in rank between the examination score and his class work exceeds one letter grade or whenever an instructor makes a request for rescoring.

Item analyses of questions used in the examinations on the basic courses prove extremely valuable in the selection of items for subsequent inclusion in freshman placement and sophomore comprehensive tests. Each year a battery of such tests is constructed, covering general education. The parts are divided so that approximately equal weight is given to scientific and verbal skill, roughly paralleling the Q and L scores for the American Council *Psychological Examination*. Repeated samples on students taking both the A.C.E. and the *Southeastern Aptitude Examinations* yields coefficients of correlation with a median value of .90.

The *Southeastern Aptitude Examinations* are constructed in April of each year, are first administered to sophomores as "comprehensives," and in the following fall are given to freshmen as placement tests. Statistical comparability for successive editions is based on the assumption that the freshman and sophomore populations are substantially the same in ability and achievement from year to year. This assumption is checked periodically by means of sampling with the same form of the A.C.E.

The framework of placement and end-of-the-year sophomore testing supplies a valuable reference for numerous studies. Relative gains over the first two years in college by fields may thus be estimated; and the results, when placed at the disposal of committees on course content, have been

found useful for revision purposes. The general or composite indices on the tests for freshmen supply "expectancies" that have been valuable in reforming grading practices in the non-basic or pre-professional curricula, in which marking has been found to be, on the average, a letter grade higher than in the basic courses, besides being for the most part extremely unreliable. Placement and sophomore test scores likewise make possible predictive studies of general ability in relation to achievement in the basic courses, where grading is relatively reliable and comparable from college to college. The part scores as well as the general score index may also be used for similar inquiries.

Coupled with the placement testings are centrally administered physical examinations. Medical staff officers and medical college seniors give their services for this purpose. The examination blank is set up for Hollerith tabulation and includes, besides quantification of clinical findings, a socio-economic scale and an index of emotional stability. Tabulation of the data makes possible, together with the "paper and pencil" testing, a variety of studies bearing on the physical and mental development of students coming from many different types of environment.

Surveys of student opinion of college work have been demonstrated as worth while in shedding light upon the effectiveness of educational practices and in comparing, from this point of view, the relative quality, difficulty, and popularity of the basic and pre-professional curricula. The setting is especially favorable to useful measures of student opinion, since approximately half of the curriculum at the junior college level is composed of basic courses common to all students and half of pre-professional or vocational courses.⁵

All examinations, forms, questionnaires and the like are prepared centrally by the photo-offset process. Collectively, the examinations of all kinds for a single year approximate

⁵"Student Opinion of College Courses, 1937 and 1940," *Examiners Office Bulletin*, September, 1940, University of Georgia Press

EXAMINERS OFFICE OF GEORGIA SYSTEM

200,000 copies. About 35 per cent of these are used outside of the University System, by colleges and high schools in the Southeast.

All data from examinations, periodic and occasional reports and studies, and general conclusions about educational policy growing out of the services and research are made available through conferences, correspondence, and formal documents to the Chancellor, the Board of Regents, and the presidents and faculties of the 15 colleges of the System. A University System Council of which the Examiner is executive secretary formulates the educational policies for the System and recommends its findings to the Chancellor and Board of Regents for action.

LEVELS OF COMPETENCE IN COUNSELING—A POST-WAR PROBLEM FOR STUDENT PER- SONNEL WORK IN SECONDARY SCHOOLS

MILTON E. HAHN
University of Minnesota

MANY THOUGHTFUL secondary school administrators are deeply concerned with the readjustment problems which will face the United States and its educational institutions after the present world conflict. The depression years of the past decade gave a fore-taste of the services which will be demanded of schools and their personnel workers in a post-war world. The inadequacy of student personnel work between 1929 and 1940 was brought sharply home to our high schools by the creation of new governmental agencies which were established to compensate for the shortcomings of public education. In many communities the first professional guidance services for youth were introduced by the National Youth Administration, the Civilian Conservation Corps, the Work Projects Administration or the Federal-State Employment Service. The work of these agencies, coupled with the relatively careful man-job analyses being made by the personnel divisions of the armed services, raises a serious question as to whether or not the public will accept traditional hit-or-miss methods in preparing post-war youth for meeting its responsibilities. School administrators face conditions which demand constructive action if their institutions are to retain the high public esteem and financial support they have enjoyed in the past.

Student personnel work is a relatively new educational configuration in secondary schools. For two decades begin-

ning about 1909, the major emphasis was upon treating individuals and their particular complex problem patterns with group methods, paralleled during the second decade by the use of tests in college and industry. The third decade of the movement was devoted to a search for tools and techniques more valid and reliable than the lecture and casual conferences between a student and a teacher. This decade contributed much to the methodology of job analysis. It was also marked by the flowing together of these two movements. With the emergence of better tools and techniques for man analysis, competent general counselors began to be trained and utilized in colleges, universities, and large secondary schools. During the 1930-40 decade the leaven of professionally trained student personnel workers spread unevenly over the country into small colleges, junior colleges, and high schools enrolling less than 500 students.

This decade also was marked by attempts to define and describe student personnel work.¹ The older term, guidance, had, because of disputes as to its nature, become more and more meaningless. Various schools of thought stretched "guidance" to mean vocational guidance only,² to be a synonym for education,³ and to cover the ordinary non-lecture activities of classroom teachers alone.⁴ There is no present definition of either guidance or personnel work which is generally accepted by all workers with youth problems. The matter of definition is of interest to us here only because it is necessary to limit the scope of our materials. Personnel work with secondary school students must be broadened in scope to include responsibilities in certain directions for out-of-school

¹The reader interested in the development of student personnel work is referred to the following sources. W. H. Cowley, "The Nature of Student Personnel Work," *The Educational Record*, April, 1936; George E. Myers, "The Nature and Scope of Personnel Work," *The Harvard Educational Review*, January, 1938; Donald G. Paterson, "The Genesis of Modern Guidance," *The Educational Record*, January, 1938.

²H. D. Kitson, "Getting Rid of a Piece of Educational Rubbish," *Teachers College Record*, XXXVI (October, 1934).

³John M. Brewer, *Education is Guidance* (New York: Macmillan, 1932).

⁴J. E. Walters, *Individualizing Education* (New York: John Wiley & Sons, 1935).

LEVELS OF COMPETENCE IN COUNSELING

youth Therefore the following working definition is offered as a frame of reference for this article.

Personnel work with youth is the marshalling, under the best obtainable professional leadership, of educational and other community resources to aid individual youth, in and out of school, to help themselves toward optimal resolutions of immediate and long-range problems in the various life-problem areas.

Because the average community is small and because the majority of workers with youth are found in the schools, the personnel program will tend to center about the secondary school for all community youth. Although this will be a typical situation, it will be necessary for the educational personnel worker employed in the educational system to refer many problems to other professional workers in the geographic or political district. At what point does the personnel worker in our schools face the necessity for referral of cases? A partial answer can be obtained through consideration of arbitrarily selected categories of personnel workers and the estimated competence for the average individual in each category relative to counseling effectiveness. Such an approach requires consideration of many variables and complicates verbal presentation. Again the writer exercises his prerogative of being arbitrary and for the sake of simplification selects the variables to be introduced. We shall consider teacher-counselors, vocational specialists, and clinical counselors as the categories of youth personnel workers. Life-problem areas will be represented by vocational problems and educational problems. Levels of case history interpretation and use of tools and techniques of the counselor are selected as the axes upon which our worker categories and life-problem areas will be considered.

Life-Problem Areas

Personnel work with individuals is necessary because they have problems which they are unable to resolve satisfactorily unaided. These problems occur in tangled patterns in which it is frequently impossible to separate one general kind of

problem from another or show clearly which is cause and which is effect. Because it is impossible adequately to verbalize a whole problem pattern, we must discuss the interrelated problems as if they were discrete phenomena. A commonly-used categorization of life-problems includes vocational, educational, personal, health, and financial adjustments. For our purposes we consider only the first two.

Vocational problems are those in which lack of adjustment is caused by poor choices of vocational field or level, no choice, or uncertainty of choice with the need for competent assurance or advice. Vocational problems may be considered in some aspects as phases of educational problems. Although vocational problems often are treated as if they were simple in nature, they are extremely complicated in many individuals. Sound vocational counseling requires that the counselor be familiar with the theory and clinical usage of the psychological concepts of abilities, aptitudes, and interests. Because of this, reliance upon untrained counselors and self-analysis has been discarded by the best practitioners. The case against these traditional methods has been stated *ad nauseum*, but these methods are still utilized in many secondary school guidance programs. A recent study by Stone⁵ presents further reasons for questioning present common treatment of vocational problems in adolescents.

Student educational problems are those caused by being thwarted in whole or in part in the attempt to proceed through a training program (usually formal) toward a goal. For many youth this goal is occupational in nature. As has already been said, vocational and educational problems are very frequently different aspects of the same general condition. Educational problems like others range from the very simple, such as a choice between afternoon or morning classes, to very complex, such as a complicated reading disability requiring special remedial work. If possible, we have placed even

⁵C. Harold Stone, "Evaluation Program in Vocational Orientation," *Studies in Higher Education*, Biennial Report of the Committee on Educational Research (Minneapolis: University of Minnesota Press, 1938-1940), pp. 131-145.

LEVELS OF COMPETENCE IN COUNSELING

greater reliance upon self-analysis for resolution of educational problems than has been true of other kinds of problems. The tragic results of past treatment of the educational problems of youth fill the literature. A pointed commentary on our educational counseling is found in the New York Regents Study.⁶

The Teachers' Level of Counseling Competence

In the student personnel programs of many secondary schools the teacher is *the* personnel worker. There are a number of reasons why this condition exists. The most important factor contributing to such programs is the concept of guidance held by so many secondary school administrators. To believe that teachers trained chiefly for classroom teaching can deal adequately with the serious problems of youth implies that the follower of this creed also believes that:

Student self-analysis has high validity and reliability.

Student problems are seldom serious.

Professional workers are not needed.

Teachers have enough free time to know each student intimately and discharge counseling responsibilities.

Tools and techniques beyond interviews and school grades and their interpretation are not worth employing or are quickly learned by classroom teachers.

The weaknesses of the teacher-counselor type of program in which many or all teachers consult on all kinds of student problems are manifold. A chief weakness of this type of program is the narrow range in which counseling competence exists.

The outline illustrates this narrow range of counseling competence. It presents crude continua of data interpretation for two problem categories—vocational and educational. It is relatively safe to assume that in neither of these continua does

⁶Francis T. Spaulding, *High School and Life*, The Regents' Inquiry (New York: McGraw-Hill, 1938).

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

the classroom teacher compare favorably with counselors trained to interpret complex interrelated data.

OUTLINE OF LEVELS OF COUNSELING

INTERPRETATION TO STUDENTS OF EDUCATIONAL PROBLEMS

Unvalidated Personal Impressions	Grades- Ratings- Student Choices	Simple Statistical Treatment of Measurement	Sophisticated Statistical Treatment of Measurement	Pattern Analysis of Individual from Synthesized Data
----------------------------------------	-------------------------------------------	------------------------------------------------------	-------------------------------------------------------------	---------------------------------------------------------------------

INTERPRETATION TO STUDENTS OF VOCATIONAL PROBLEMS

Unvalidated Student Choices	Occupational Application of Specific Sub- ject Matter	Relation of School Subjects to World of Work	Occupational Infor- mation	Simple Statistical Treatment of Meas- urement	Sophisticated Statistical Treatment of Meas- urement	Pattern Analysis of Individual from Synthesized Data
-----------------------------------	----------------------------------------------------------------	----------------------------------------------------------	----------------------------------	-----------------------------------------------------------	------------------------------------------------------------------	---------------------------------------------------------------------

Teacher Counseling and Educational Problems.—In the area of educational problems the average teacher works for the most part with unvalidated student statements or with equally invalid personal impressions. Many teachers interpret their grades and ratings to youths seeking counsel. A few teachers have become skilled to an extent that they can interpret various kinds of data in terms of simple statistical concepts. A very few are competent to interpret complex, related data, dependent for its meaning upon great statistical sophistication. A rare individual can utilize job and man analyses in such a way that proper interpretation is supplied the counselee. Ineffective educational counseling by secondary school teachers is not a matter of speculation or assumption. Eckert and Marshall⁷ state that more than three of five high-school students in New York State leave school before graduation. Many of those leaving school do so because of inability to meet the demands of the curricula in which they attempt to compete. Many of these students could profit from courses of study different from the ones in which they failed.

⁷Ruth E. Eckert and Thomas O. Marshall, *When Youth Leave School, The Regents' Inquiry* (New York: McGraw-Hill, 1938), pp. 48-49.

LEVELS OF COMPETENCE IN COUNSELING

Edgerton and Toops⁸ estimate that only 34 per cent of 1,958 students included in a survey of Ohio college students had records indicating ability for eventual college graduation. Many of these students must have been advised by high-school teachers and administrators to attempt higher education at the college level. The literature of secondary school and college mortality points clearly to a large amount of poor advising about educational problems. Williamson⁹ offers a satisfactory summarization of educational counseling by teachers when he says:

Fewer students would select inappropriate courses if reliable statements of requirements of the wide variety of occupations and professions open to high-school graduates were available. . . The counseling use of such information would enable more students to prepare for appropriate occupational goals.

Many scholastic failures could be avoided if administrators and teachers would establish comparable and valid standards, so that students and counselors could better judge of future success in a course by past achievements in a related course.

Teacher Counseling and Vocational Problems.—Because educational and vocational problems of youth are so closely related, much that has been said of teacher counseling and educational problems can also be said of vocational problems. The effects of poor vocational counseling upon the boy or girl are often more serious than those of poor educational counseling. If a student takes a course for which he is not suited, adjustments can be made through failure or change of course. These adjustments do not require long time spans. Poor vocational counseling can result in situations where the unadjusted individual may be forced by circumstances to spend long periods in which change is difficult or impossible. One has only to inspect the occupational choices of high-school

⁸Harold A. Edgerton and Herbert A. Toops, "Academic Progress," *Contributions in Administration* 1 (Columbus, Ohio: Ohio State University Press, 1929), p. 136.

⁹E. G. Williamson, *How to Counsel Students* (New York: McGraw-Hill, 1939), pp. 260-261.

seniors to realize the unrealistic thinking which can supply fertile ground for poor advising. Stone¹⁰ demonstrates that in one freshman group at the University of Minnesota only 40 per cent of the students had occupational goals which were judged valid by competent clinical psychologists. A goodly proportion of these students came from high schools which have prided themselves upon their teacher counseling programs over a long number of years. Students from these schools had no better choices than those from schools which made no pretensions to vocational counseling. Stone's study indicates that counseling by professionally trained, clinical counselors reduces the number of poor vocational choices and increases the number of good choices. The gains are statistically significant. Williamson and Bordin¹¹ found that control-group (uncounseled) college students achieved a vocational adjustment judged to be satisfactory by themselves and the evaluating judges in 68 per cent of the cases. On the other hand, such an adjustment was achieved by 81 per cent of the cases in the experimental group (counseled by clinical counselors). Satisfactory adjustment was not made by 27 per cent of the control group and 15 per cent of the experimental group. These differences are statistically significant.

Large numbers of high schools depend upon classroom teaching of occupational information to resolve vocational problems of students. This faith in "talking at" students has little to recommend it. Many of the studies which show advantage to classroom group-counseling do so only in terms of gains in amount of occupational information. No one has produced evidence that students with the greatest amount of occupational information make the best vocational choices. It is obvious that a student who takes a course in any field of knowledge *should* know more about it than the student who has not had the same or similar courses.

Many useful tools and techniques of counseling have been

¹⁰C Harold Stone, *op cit*

¹¹E G. Williamson and E. S. Bordin, "Evaluating Counseling by Means of a Control-group Experiment," *School and Society*, LII (1940), 434-440

originated or improved in the past decade. Teacher-counselors are seldom trained to collect and collate data originating from these instruments and methods. They cannot be expected to be both teachers and applied psychologists. If teachers become clinical counselors, they no longer are classroom teachers. We need not expect adequate counseling in regard to the vocational problems of youth until our schools make use of persons other than classroom teachers to assume and discharge at least supervisory responsibilities for counseling. As will be stated later, this does not mean that each small school unit must or should have a professionally trained counselor or close up shop.

Referral to the outline on page 248 indicates that relatively few teachers are so trained that they can interpret data to students adequately if such data involve more than the presentation of information of a simple nature. Sound job analysis by teachers offers serious difficulties. Valid man analysis is beyond the ken of the average teacher.

The Vocational Specialist's Level of Counseling Competence

The vocational specialist appeared on the personnel work scene in the second decade of the developing secondary school personnel work movement. A growing public sense of need to meet the pressing problems of youth forced educators to take cognizance of these problems. The movement was first directed toward emphasis upon "things to be done" rather than toward "men and women who do things"—job analysis, not man analysis. This trend was clearly reflected in the proposed qualifications for counselors which appeared in the literature of that period. Myers,¹² for example, wrote:

It is well to remind ourselves, however, that among the qualifications, aside from special training, which those who select counselors often emphasize are: (1) a personality which attracts and gets on well with adolescents; (2) sufficient maturity to command the respect of pupils and fellow

¹²George E. Myers, "A Training Program for Counselors," *Vocational Guidance Magazine*, p. 315 V (1927).

teachers; (3) *at least as good a general education as is possessed by the average high school teacher, usually represented by graduation from a college in good standing*, (4) successful experience as a teacher; and (5) *preferably, some business or industrial experience.* (Italics not in original.)

The Committee on Standard Certification of Vocational Counselors, a committee of the National Vocational Guidance Association, advocated the following college courses for permanent certification:¹³

1. General courses—the usual courses required of candidates for teaching certificates: educational psychology, principles of teaching, educational measurements, sociology, economics.

2. Related courses — *principles and problems of vocational education, industrial history, labor problems.*

3. Guidance courses — *principles and problems of vocational guidance, analysis of vocational activities, methods of imparting occupational information, psychological tests in guidance, counseling the individual, placement and follow-up, and field work in guidance.*

Inspection of these training programs indicates clearly the emphasis placed upon the job and the worker's relationships to it. Adequate tools and techniques had not as yet been discovered to analyze and treat the individual as some one to whom a particular set of duties could be fitted. Practice was to treat the job as a set of duties to which a man or woman must be fitted. Vocational specialists dealt with *vocational problems*, i e, job specifications, occupational trends, labor problems, how to get a job, placement, and follow-up. Vocational aspects of total adjustment were considered so important that other aspects of human adjustment were often overlooked or left to be treated by specialists not yet found in secondary schools.

The unfortunate feature of the era of vocational specialists is not that personnel work passed through the stage, but rather that the stage has persisted. Too many vocational

¹³Leonard V. Koos, and Grayson N. Kefauver, *Guidance in Secondary Schools* (New York: The Macmillan Company, 1932), pp. 569-673.

LEVELS OF COMPETENCE IN COUNSELING

specialists continue to think of personnel work as job description and placement long after educational leaders have relegated their particular contribution to an important but subsidiary position in the field.

Vocational interests are no longer what boys and girls say they want to do (usually stated as a job label). Vocational interests are analyzed today by considering youths' claims in the light of observed behavior over a period of time and the leads furnished by various technical psychological measuring instruments. Selection of a career is no longer in terms of whether or not an individual can do a job. Rather the question is raised regarding what family¹⁴ of jobs and at what level within this family the optimal vocational adjustment will tend to occur. We are not so prone to encourage an adolescent in a secondary school to be a doctor of medicine. We tend to direct to "scientific fields at the professional level."

Evidence of failure to meet the vocational problems of youth through the services of vocational specialists is abundant in the literature. Anderson¹⁵ made a strong case for psychiatric services in industry. The conditions cited by Anderson raise questions as to the ways in which the men and women studied were guided to their occupational niches. Fisher and Hanna¹⁶ also contributed evidence that an alarming number of workers were not being directed or helped into suitable careers. There is little evidence to show that the worker who had the help of the vocational specialist made significantly better occupational choices than his non-counseled brother.

Reference to the outline leads one to suspect that the vocational specialist has been handicapped in his work by his general inability to deal with man analysis. His contribution to personnel work, however, has not been small. The shift

¹⁴Donald G. Paterson, Clayton d'A. Geiken, and M. E. Hahn, *Minnesota Occupational Rating Scales* (Chicago: Science Research Associates, 1941).

¹⁵V. V. Anderson, *Psychiatry in Industry* (New York: Harper and Brothers, 1929).

¹⁶V. E. Fisher and Joseph V. Hanna, *The Dissatisfied Worker* (New York: Macmillan, 1931).

to emphasis upon men rather than jobs was hastened appreciably by his efforts. Nevertheless, general responsibility for counseling of youth can hardly be delegated to the vocational specialist. Except for a minor specialty, he is no more competent than the teacher-counselor.

The Clinical Counselor's Level of Counseling Competence

The clinical counselor is a highly trained, widely experienced, applied psychologist. His training has been directed primarily to an understanding of people both as unique individuals and as members of various groups. He is a specialist in one or more areas of human problems. At the same time he is a generalist, albeit with knowledge of his limitations. The place of the clinical counselor in the field of personnel work is well stated by Williamson¹⁷ when he says:

While clinical counseling is only one of several specialized fields dealing with personal problems, we maintain, however, that it is the basic type of personnel work with individual students and serves to coordinate and focus the findings and efforts of other types of workers.

Paterson, Schneidler, and Williamson¹⁸ contend that persons training to be clinical counselors should complete the Master's degree in psychometrics or its equivalent. Further they consider the Ph.D. degree or its equivalent in technological psychology as desirable. The counselor so trained should be a master of the tools and techniques that fill the competent counselor's kit. The clinical counselor should be competent in the full range of data interpretation found in the outline.

It is not our present purpose to analyze the competency and functions of the counselor. It is safe to assume that such a one is, at the present time, our best trained personnel worker

¹⁷E. G. Williamson, *How to Counsel Students* (New York: McGraw-Hill, 1939), p. 36.

¹⁸Donald G. Paterson, G. Schneidler, and E. G. Williamson, *Student Guidance Techniques* (New York: McGraw-Hill, 1938), pp. 302-303.

with the vocational-educational problems of youth. Evidence has been marshalled which indicates that counselors at this high level of competence do achieve appreciably better outcomes of counseling than is true of other individuals working with vocational-educational problems of youth. It is our purpose to urge that we make use of these people even in small school systems which cannot add them to their full-time staffs.

The average American high school is small. Sound school-community personnel programs must include the pooling of resources in order that these small schools can have many kinds of services which they could not afford alone. Administrators in small secondary schools will find that complete youth personnel programs are beyond them when they consider only the community which they serve. When consideration is given to the combined resources of five, six, or seven schools, there are practical solutions to the problem. Sharing the services of clinical counselors is such a solution. When an administrator faces an in-service training program for teacher-counselors with no professional assistance, he is involved in difficulties. When he faces in-service training of teacher-counselors as part of a county or district project in charge of a competent instructor, many of these difficulties disappear.

Few small communities can supply enough personnel work with youth to occupy the full time of a clinical counselor who concentrates upon vocational-educational problems. Part-time aid from such a counselor will, in many instances, be all that is needed to develop the school-community personnel program. A qualified counselor can discharge personnel functions in several schools. For example, research on problems common to several schools may be conducted almost as easily as for a single institution. In-service training in a district may be little more difficult than in a single institution. Counseling of difficult cases in a district may involve no greater case load than would be true in a single large institution or community. Development of several sound school-community programs at one time in cooperation with other personnel agencies is not an unreasonable task.

We have had time to discover the gross errors in assigning major counseling responsibilities to teachers and vocational specialists. We have not yet had time to correct these errors. Personnel work with youth in the post-war period will go forward, although there is no guarantee that the public schools will retain the golden opportunity they have had to develop the field. If secondary school administrators will forget tradition and face the tasks ahead realistically, if they will turn from subject matter and "things for people to do," if they will make use of the best personnel workers, they may yet retrieve the losses in public support and esteem which they suffered in the depression and war years. Much depends upon the level of counseling effectiveness which the schools achieve.

A STUDY OF SOME LOCAL FACTORS AFFECTING STUDENTS' SCORES ON THE MINNESOTA PERSONALITY SCALE

BETTY M. HORNE AND W. C. McCALL
University of South Carolina

A FRESHMAN TESTING program including tests of general scholastic aptitude as well as tests of achievement and aptitude in specific subject fields has become common practice in our colleges and universities. The results of these tests are customarily used in counseling with the student on academic and vocational problems, and in predicting the student's probable academic success.

Many institutions also include in their testing program instruments which are designed to measure the student's vocational interests and aptitudes and scales designed to evaluate his personality adjustment. Both the felt need for information of this nature as an aid to more effective guidance procedures and the increased reliability and validity of recent measuring instruments have undoubtedly influenced this trend.

However, the effect of local factors on all test scores and more particularly on tests of personality has long been recognized. The present study reports a systematic attempt to analyze these factors in a specific situation. In September, 1941, the Minnesota Personality Scale was added to the regular list of freshman tests at the University of South Carolina. The test was thus administered to 241 freshman men and 144 freshman women.

Two different forms of the tests have been published, one for men and one for women. The total scale consists of five sub-scales measuring, respectively, Morale, Social Adjustment,

Family Relations, Emotionality, and Economic Conservatism. The authors of the test, Darley and McNamara, describe these sub-scales as follows:

Part I—Morale: High scores are indicative of belief in society's institutions and future possibilities. Low scores usually indicate cynicism or lack of hope in the future.

Part II—Social Adjustment: High scores tend to be characteristic of the gregarious, socially mature individual in relations with other people. Low scores are characteristic of the socially inept or undersocialized individual.

Part III—Family Relations: High scores usually signify friendly and healthy parent-child relations. Low scores suggest conflicts or maladjustments in parent-child relations.

Part IV—Emotionality: High scores are representative of emotionally stable and self-possessed individuals. Low scores may result from anxiety states or over-reactive tendencies.

Part V—Economic Conservatism: High scores indicate conservative economic attitudes. Low scores reveal a tendency toward liberal or radical points of view on current economic and industrial problems.

As is customary at the University of South Carolina, norms for the local population were set up and have been used in rating all students. A comparison of these norms with those published for the University of Minnesota population on which the test was standardized comprise the first results of this study, and are presented in Table 1.

Comparison of Minnesota and South Carolina Scores

Since the Minnesota norms are stated in terms of percentile values, it was necessary to base all statistical calculations on these values. Accordingly, the critical ratios are expressed in terms of the difference between fiftieth percentile points divided by the P.E. of the difference between these medians and must equal 4.0 or more in order to be statistically significant. Since there are different forms of the test for men and women, the critical ratios for the two sexes have been calculated separately.

LOCAL FACTORS AFFECTING SCORES

TABLE 1

COMPARISON OF SCORES OF MINNESOTA STUDENTS AND SOUTH CAROLINA STUDENTS ON THE MINNESOTA PERSONALITY SCALE

Median Raw Score	Sub-Scales				
	I	II	III	IV	V
Minn. men (N = 1083)	167	224	138	159	106
S. C. men (N = 241)	172	230	149	163	105
Critical ratio*	6.5	3.3**	11.3	3.6	1.9
Minn. women (N = 888)	173	228	149	168	104
S. C. women (N = 144)	178	237	158	170	104
Critical ratio ⁺	5.2	4.7**	7.8	1.1	0

*Critical ratio = Difference between medians divided by probable error of this difference

**Critical ratio of difference between Minnesota median and South Carolina mean for women = 2.1, for men = 2.2.

The table indicates two areas in which the South Carolina students appear to be better adjusted than the Minnesota students, Morale and Family Relations. The critical ratios for these scales are significant for both men and women. Data to be presented below indicate that the latter difference, that in Family Relations, is probably related to the relatively smaller number of large centers of population in South Carolina than in Minnesota.

The explanation for the difference in general morale is less apparent. The items in this sub-scale may very roughly be divided into three main groups, namely, questions concerning faith in the honesty and adequacy of our legal system, questions dealing with faith in the value and methods of our educational system, and faith in the possibilities which the future holds for the individual. It must be emphasized that this division is both arbitrary and rough, and since no data were available for an item analysis, no comparisons within this sub-scale are possible. A generalized statement based on the authors' description of the area of personality adjustment measured by this scale would indicate that the South Carolina students had significantly more faith in society and

its institutions and in their own future than did the Minnesota students.

Scores of entering South Carolina freshmen on the *Test of General Proficiency in the Field of the Social Studies* of the *Cooperative General Achievement Tests* indicate that their acquaintance with the strengths and weaknesses of our American social institutions is more limited than that of the 6296 freshmen on whom the test was standardized. The South Carolina mean fell at the thirty-fifth percentile point for the standardization group. It is possible that this lack of acquaintance has tended to promote an uncritical acceptance of these institutions. It is not improbable that this difference, also, is related to the factor of population distribution, although the subsequent data do not strongly suggest such a conclusion.

The scores on Scale II would suggest that the South Carolina students are more interested in and better adjusted to social group life. The nature of the questions on this scale strongly suggests that it measures a factor closely resembling the usual definitions of introversion-extroversion. Although the foregoing statement may be interpreted as lending support to the tradition of hospitality and sociability of the Southern home, the statistics also offer an alternative explanation.

The distributions of scores on this particular sub-scale seem to be somewhat skewed, since the median women's score is 5 points higher than the mean and the median men's score is 2 points higher than the mean. The discrepancy of 5 points in the women's distribution is the greatest difference between median and mean in any of the 10 distributions, the two differences on Scale III being 2.5 points, and all others being 2 points or less. It is the only instance in which such discrepancy affects the significance of the difference between the scores of Minnesota and South Carolina students. As the footnote to the table indicates, the difference between the Minnesota median and the South Carolina mean for women on Scale II, divided by the probable error of the difference between the two medians, is only 2.1, a critical ratio which

LOCAL FACTORS AFFECTING SCORES

is not significant; the corresponding ratio for men is 2.2. The reader may choose the explanation of these data which seems to him most logical and acceptable

Relation of Scores to Size of Home Town

We have referred above to an analysis of the data from South Carolina students based on size of population centers. The University is located in Columbia, the capital of the state and a city of about 75,000. Thirty-five to forty per cent of the students are from Columbia. The comparison of adjustment scores for students from population centers of different sizes was originally suggested by a clinical observation that there seemed to be a disproportionate number of students from Columbia who showed one single low score on Scale III, Family Relations.

The complete set of data is presented in Table 2. The students were divided into three groups, according to their home residence and the size of the town in which they attended high school. Class A includes all students who resided in and attended high school in cities of 25,000 and over; Class B includes those who resided in and attended high school in towns of 2500 to 25,000; and Class C those

TABLE 2

COMPARISON OF SCORES OF SOUTH CAROLINA STUDENTS FROM TOWNS
OF CLASS A, CLASS B, AND CLASS C*

Mean T-score value	Sub-Scales				
	I	II	III	IV	V
Class A (N = 183) . . .	49	47	44	48	47
Class C (N = 80) . . .	51	50	47	48	49
Class B (N = 91) . . .	52	56	55	54	52
Critical ratios**					
Class A vs. Class B . . .	1.0	3.6	4.3	2.3	2.1
Class B vs. Class C . . .	0.3	2.2	2.8	2.1	1.1
Class A vs. Class C . . .	0.6	1.1	0.9	0.2	0.6

*Class A—Towns with a population of 25,000 and over.

Class B—Towns with population of 2500 to 25,000.

Class C—Towns with population of less than 2500 and rural districts.

**Critical ratio = Difference between means divided by the standard error of this difference.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

who attended high school in a town of 2500 or less and resided in a town of that size or gave a rural address.

It should be noted that Class B was originally divided into groups from towns of 2500 to 10,000 and from towns of 10,000 to 25,000. Differences in average score between these two groups were generally small, the group from towns of 10,000 to 25,000 being in most instances somewhat lower. However, the number of students falling in the group from 10,000 to 25,000 was so small as to make any real conclusions impossible. The present Class B is composed of 68 students from towns of 2500 to 10,000 and 23 students from towns of 10,000 to 25,000.

TABLE 3
MEASURES OF VARIABILITY FOR AVERAGES PRESENTED IN
TABLES 1 AND 2

Median Raw Score \pm P.E.	Sub-Scales				
	I	II	III	IV	V
Minn. men (N=1083) . . .	167 \pm 9	224 \pm 21.5	138 \pm 12.5	159 \pm 12	106 \pm 8
S. C. men (N=241) . . .	172 \pm 8.5	230 \pm 20	149 \pm 10.5	163 \pm 12.5	105 \pm 5.5
Minn. women (N=888) . . .	173 \pm 9	228 \pm 18	149 \pm 14	168 \pm 16	104 \pm 6
S. C. women (N=144) . . .	178 \pm 8.5	237 \pm 17	158 \pm 9.5	170 \pm 16.5	104 \pm 5
Mean T-score \pm S.D.					
Class A (N=183) .	49 \pm 21	47 \pm 22	44 \pm 21	48 \pm 20	47 \pm 19
Class C (N=80) .	51 \pm 22	50 \pm 19	47 \pm 20	48 \pm 19	49 \pm 20
Class B (N=91) .	52 \pm 18	56 \pm 19	55 \pm 19	54 \pm 19	52 \pm 16

An explanation of certain statistical techniques used in the calculations of Table 2 is necessary. All individual scores for South Carolina freshmen were translated into T-score values determined from the means and standard deviations of the distributions of raw scores. The scores presented are, therefore, average T-score values, not average raw score values. The use of these T-score values provides a basis on which scores for men and women may be combined for statistical treatment. The T-scores for each sex are based on the scores of that sex, but a given T-score value for men and women is considered comparable.

For convenience in reading, Class C is shown between Class A and Class B in Table 2. The evidence provided by

the table is both striking and self-explanatory. On every scale the average scores of students gradually increased from Class A through Class C to Class B. At least for South Carolina high school students the environment most conducive to personality adjustment seems to be a Class B community; that is, a town from 2500 to 25,000. Very small towns or rural districts appear somewhat more favorable than metropolitan centers.

Critical ratios presented in this table are based on the difference between means divided by the standard error of this difference, and are thus significant at 3.0. The differences between Class A and Class B students on Scales II and III, Social Adjustment and Family Relations, meet this criterion; and the differences between Class B and Class C show 98.6 and 99.7 chances in 100, respectively, of a true difference on these same scales.

It seems safe to conclude, therefore, that towns of medium size provide a significantly better background for the development of social maturity and extroverted social relationships than either a city or a very small community. Speculation concerning the complex factors operating to produce these differences would be interesting but highly subjective.

At least one factor affecting the home and family adjustments would seem, however, to be less complicated. One large group of questions contained in this sub-scale relates to possible maladjustments arising from the young person's efforts to establish his social and personal independence. It seems logical to suppose that these areas of family relationship are subject to greater strain in either a metropolitan center or a small town than in a medium-sized urban community. The "temptations" of city life have been discussed perhaps far too much in our recent sociological literature, but the opportunity and motivation toward social independence offered by the recreational, social, and even school-sponsored activities of a city are too obvious to require elaboration. Any effort of the parents to counteract these influences will almost inevitably lead to family conflict.

In the rural home, on the other hand, it seems quite probable that the source of conflict lies in the moral conservatism which is generally assumed to be characteristic of farm parents. We are here assuming, of course, that towns under 2500 resemble rural areas in their mores and attitudes, an assumption which is probably justified. If this is the case, rural and small town students undoubtedly find themselves in conflict with their parents over proposed activities which would receive no frown of disapproval from the city parent.

The authors are aware that the foregoing paragraphs of interpretation involve several assumptions with which the reader may disagree. The data are clear-cut in their exposition of the facts; the interpretation of these facts must of necessity be somewhat subjective.

On a third sub-scale, number IV, Emotionality, the data indicate differences approaching statistical significance between students of Class A and Class B, and between Class B and Class C. These differences are probably related to the correlations presented in Table 4. Since the only correlations in this table above .50 are those between Scales II and IV and Scales III and IV, at least some of the factors which influence Scales II and III must influence Scale IV in a similar direction. It appears probable that the differences on Scale IV are dependent on these relationships to some extent.

Inter-Relationships Between Scales

The data of Table 4 are of special interest on two points. The similarities of the correlations based on the scores of Minnesota students and of South Carolina students are striking. The critical ratios of the differences of these various correlations were calculated, and of the 20 comparisons thus made, only one critical ratio, that between Scales II and IV for men, was above 2.0, and this one did not reach significance. As the table indicates, the range of average correlations for the four groups studied is only .02.

The second point of interest concerns the relationships of Scale IV to other scales in the test. This sub-scale meas-

LOCAL FACTORS AFFECTING SCORES

TABLE 4

INTER-CORRELATIONS OF THE FIVE SUB-SCALES OF THE
MINNESOTA PERSONALITY SCALE

Correlation	Minnesota Men (N=577)	S. Carolina Men (N=241)	Minnesota Women (N=557)	S. Carolina Women (N=144)
Scale I with II.41	.37	.36	.31
I with III.26	.33	.34	.26
I with IV38	.36	.38	.34
I with V21	.22	.18	.18
Scale II with III25	.36	.26	.32
II with IV53	.39	.48	.51
II with V17	.19	.13	.11
Scale III with IV52	.56	.54	.58
III with V24	.25	.16	.17
Scale IV with V21	.13	.15	.28
Average32	.32	.30	.31

ures emotional stability, and includes many questions usually found in an inventory of neurotic traits. The fact that it reflects and is reflected in other areas of personal adjustment is, therefore, quite in harmony with repeated observations of clinical psychologists. As has been pointed out, the only correlations above .50 in the table involve this sub-scale. The average inter-correlation of this scale with all other scales, for all groups involved, is .40. The average inter-correlations of the other scales, exclusive of their correlation with Scale IV, are .29, .27, .27, and .18, respectively, for Scales I, II, III, and V.

Summary and Conclusions

The Minnesota Personality Scales for men and women have been administered to 241 freshman men and 144 freshman women at the University of South Carolina. The scores of these students have been compared with the norms data published for the scale, based on the scores of 1,083 fresh-

man men and 888 freshman women at the University of Minnesota. The total scale consists of five sub-scales which measure Morale, Social Adjustment, Family Relations, Emotionality, and Economic Conservatism.

The data herein presented support the following conclusions:

(1) South Carolina students, both men and women, obtained scores indicating a significantly better adjustment in Morale and in Family Relations than those of the Minnesota students. There is some evidence that the South Carolina students are superior in Social Adjustment, though this conclusion is not clearly substantiated.

(2) South Carolina students from towns of 2500 to 25,000 population (Class B) appear somewhat better adjusted on all scales than students of towns of 2500 or less (Class C), and the latter slightly better adjusted than students from cities of 25,000 and over (Class A). These differences reach significance between Class A and Class B in Social Adjustment and in Family Relations, and approach significance between Class B and Class C on the same scales. They approach significance between Class A and Class B and between Class B and Class C in Emotionality.

(3) Inter-correlations between the several sub-scales based on data from Minnesota students and from South Carolina students are strikingly similar. The scale measuring Emotionality shows the only inter-correlations above .50, and shows a higher average inter-correlation than any other sub-scale. Correlations between Emotionality and Social Adjustment and between Emotionality and Family Relations are above .50.

REFERENCES

1. Dailey, John G., and McNamara, Walter J. *Minnesota Personality Scale, Manual of Directions*. New York: The Psychological Corporation, 1941.
2. Darley, John G., and McNamara, Walter J. *Minnesota Personality Scale (For Men)*. New York: The Psychological Corporation, 1941.
3. Darley, John G., and McNamara, Walter J. *Minnesota Personality Scale (For Women)*. New York: The Psychological Corporation, 1941.
4. Willis, Mary, et al. *Cooperative General Achievement Tests, Number I A Test of General Proficiency in the Field of the Social Studies, Form QR*. New York: The Cooperative Test Service, 1940.

THE PLACE OF APTITUDE TESTING IN THE PUBLIC SCHOOLS

DONALD E SUPER
Clark University

IN THE PRACTICE of aptitude testing, three basic assumptions are important. These assumptions have been so well established by research in the psychological laboratories, in the schools, and in industry that they are now generally taken for granted and need little justification.

One assumption is that individuals differ in the extent to which they possess any given aptitude, some being well endowed with the aptitude, let us say, to sing, others having little aptitude for vocal music, and most of us being potentially only mediocre singers.

The second assumption is that there are a number of special aptitudes, such as aptitude for musical expression, aptitude for mechanical work, aptitude for visualizing the relations of objects in space, scholastic aptitude, manual dexterity, and aesthetic judgment.

The third assumption is that there are important differences in the amounts of these various aptitudes possessed by a given individual

Dr. Walter Dill Scott, pioneer industrial psychologist and until recently president of Northwestern University, has an interesting story illustrating this point. According to personnel data compiled by him, the most successful salesman in a wholesale food company was also its least intelligent salesman. Unable to reconcile these two items of information, Dr. Scott investigated further, found that this was indeed the case, and sought an explanation. He found that the salesman would go into a delicatessen, let us say, and chat with the owner and his wife. The conversation generally dealt with family and neighborhood affairs, about which the salesman kept posted. Finally he would get around to pickles and other items of

business. Then another man would enter the picture, a second salesman employed to work with him, who discussed prices, took orders, filled out blanks, and performed other clerical tasks which the star salesman could not handle. It actually paid the company to employ two men to do one man's work! Such extreme variations of abilities in one individual are the exception, as Dr. Terman demonstrated in his "Genetic Studies of Genius," but the extreme case illustrates a less extreme tendency toward stabilization of aptitudes and abilities within individuals.

These three assumptions have provided us with a basic philosophy of education and of guidance, together with a working program for the schools. Recognizing the potential worth of each individual, it becomes incumbent upon us, as members of a democracy, to provide for individual differences in the children with whom we work. It is also important, if we are to make our democratic system effective, to study the individual differences in our pupils and to help them understand their own abilities and interests, in order that they may choose wisely from among the various educational offerings provided. It is not enough to develop differentiated curricula, as will shortly be demonstrated, unless we also provide the means of making wise choices of curricula. It is at this point, of course, that aptitude tests enter the picture.

Before proceeding to discuss these last in some detail, let us dwell briefly on each of these two aspects of the working program of a democratic educational system, curricular differentiation and individual analysis, imposed upon us by our recognition of individual differences, special abilities, and trait differences.

The history of American secondary education is in effect the history of a long drawn-out and not altogether conscious attempt to provide for individual differences. The colonial Latin Grammar School existed to provide pre-professional education, to prepare for college boys who were to enter the learned professions. It was largely supplanted by the Acad-

emy, the purpose of which was to add two new types of education for two new types of pupils. It offered scientific and commercial training for those who were planning to enter technical occupations and the field of business, in addition to academic courses. The public high school entered the picture in the last century in order more effectively to provide these same types of education. Its purpose was to offer pre-professional and pre-commercial training, as an analysis of its subjects and of the then current vocational conditions would show. In more recent years the Industrial Arts course and the Trade School have been developed to meet the needs of those who are likely to enter the skilled trades. Some of the recent evaluative studies of public education, such as the Regents' Inquiry in New York State, now advocate the development of a fourth type of secondary education, a high school with a curriculum designed to prepare youth for work in the semi-skilled trades and for the patterns of living typical of those employed at that level. A few schools already provide such courses.

If we analyze the trends so briefly described above, we see at once the increasing differentiation of our educational offerings as a result of the recognition of individual differences and the demand for appropriate curricula.

One might expect, once a reasonable variety of educational offerings is provided, that the distributive mechanism of a democratic educational system would function smoothly. Pupils and their parents could look over the offerings and choose appropriate courses, especially if given the benefit of the advice of teachers and principals who are familiar with both the children and the courses. The practices of many schools have been based on this assumption. Recent years, however, have shown that this is unwarranted, for numerous thorough studies have indicated large numbers of young persons obtain an education of a type not suited to their vocational prospects.

This statement should be illustrated with concrete facts, for such claims are all too frequently made without adequate foundation. Two-thirds of our youth of high school age

attend high school, that is, are in schools which prepare for professional, commercial, or skilled employment. But only one-fourth of our employed adult population is actually engaged in occupations of these types. This means that five-twelfths, or approximately one-half, of the young people whom we are attempting to educate are actually being prepared for vocations and for ways of life which they will not enter. To put this in another way, our young people now tend to get an education planned in terms of the upper half of the occupational and social scale, whereas most of them enter and remain in occupations in the lower half of the scale. Surely no further proof is needed that young people need vocational guidance, that is, help in understanding and acting on their own abilities, interests, and opportunities.

Given several different types of secondary school curricula, and granted the need to help students decide which type of curriculum to pursue, we must then devise methods of individual analysis which will assist them in making curricular and consequent vocational choices.

Various methods suggest themselves. We may examine a student's marks in the different subjects which he has studied and find out in which types he has done the best work. But this approach has at least two important limitations: the courses which he has taken have been limited in variety and in number, and teachers' marks are frequently unreliable and invalid indices of the quality of the work done. Useful as such data are in understanding a pupil, they cannot in and of themselves suffice for the task at hand.

We may keep cumulative records in which are noted not only the pupil's marks, but his extra-curricular activities, his relations with his fellows, his special interests and out-of-school activities. These can, as we know from their wide use, be very helpful in understanding a child. But, again, the experiences are likely to be limited in variety (a defect which can be at least partly overcome) and the evaluations made of these experiences are necessarily subjective. They fre-

quently do not permit comparisons with other persons, and their real significance for curriculum and for vocations is too often not clear.

Perhaps a digression is desirable to illustrate this last point, namely, the doubtful nature of some of the relationships which we think we see between hobbies and school subjects or vocations. It is widely assumed among philatelists that as a result of collecting stamps they learn a good deal more concerning history, geography, and related subjects than they otherwise would. To check up on this assumption, a series of studies of adolescent and adult stamp collectors were made. They were given tests of intelligence, of achievement in the social studies, and of technical philatelic matters. The same tests were given to a control group of non-philatelists. We found that the adult stamp collectors had learned a great deal more about stamp collecting, a little more about strictly factual aspects of geography (such as names of capitals), and nothing more about significant social problems. The adolescent stamp collectors learned nothing but the technology of philately from their hobby. These and other studies suggest that information concerning a pupil's activities must be used with caution as an index of aptitude, of achievement, or of interest in supposedly related fields.

It should be clear that aptitude tests are needed as a supplement to these other not too effective methods of analyzing human abilities, interests, and achievements. They are needed because, when well constructed and wisely used, they are objective, because they make possible comparisons between people, and because their curricular and vocational significance can be established with relative ease. These three concepts of quantification, reliability, and validity are now so generally familiar that they need not be elaborated.

We may ask next: What aptitude tests should one use in a school, and at what age should they be used? Before answering that question we must find the answer to another: What do we want to measure, and when must we measure it?

To reply to these questions in general terms at first, we want to measure those characteristics which are important in success at a given stage of a child's educational or vocational career some time before he enters that stage, in order that he may plan for it with wisdom. This means that different types of traits and abilities may well be measured at different ages and stages, as life's decisions make those data desirable.

What are these stages when decisions are being made? One of the first, obviously, is when the pupil leaves the elementary or junior high school to enter high school. Another is when he leaves high school to enter a vocation or a college. Still another is when he leaves college to enter an occupation. If our schools are in fact pre-vocational, the data needed at one stage are substantially the same as those needed at another.

When a student leaves the lower school to enter high school, he has to make decisions concerning the type of high school and the type of curriculum, concerning the specific course within the curriculum, and, since the curricula are in a sense pre-vocational, concerning the general family of occupations which he wishes to enter. These decisions must be based on the abilities of the pupil as they relate to the requirements of the courses. Tests should therefore be selected so that they will tap the various aptitudes, interests, and achievements which make for success and satisfaction in those courses.

To profit from the college preparatory course a pupil should have high average or superior mental ability, for much of the content of the course is abstract; an extensive vocabulary, since the subject matter is contained in books and since its exercises are generally verbal; superior reading speed and comprehension, for the same reasons; ability to work with numbers, since numerical symbolism and manipulation are important in many subjects, especially for the future technologist; interest in the why, how, and whence of things and of ideas, since of such is the content of most academic subjects and the basis for most of the professions. To the prospective college

preparatory or academic student one would, therefore, want to give tests of scholastic aptitude, vocabulary, reading, mathematics and other academic subjects, and interests

Are such tests available for people as young as 14 or 15? Intelligence or scholastic aptitude tests have, of course, long been in use at all ages. Achievement tests in the tool subjects are equally well standardized and validated. Both can be given by teachers with a minimum of training in test techniques and can be scored for relatively little money. Interest tests are not so well developed at this level, but there are at least two, and probably three or four, which can be given to early adolescents with some confidence if the results are to be used by competent people. Scoring may run into more money, but the cost is not necessarily prohibitive.

Singularly little attention has been given in most localities to the qualities needed for success in the commercial course, although this has to some extent been remedied in more recent years. Again, intelligence or scholastic aptitude plays a part, although it need not be present in the same degree as in the college preparatory course. A good deal of subsequent disappointment would probably be avoided if most pupils with I.Q.'s of less than 105 or 110 could be motivated to choose courses other than the academic, and if most of those with I.Q.'s of less than 95 or 100 could similarly be guided (but not coerced) into courses other than the commercial. A moderately good vocabulary, reading ability, and mathematical achievement are required here too, although the minimum requirement is somewhat lower than that of the academic course. The interests of commercial employees are different from those of professional people not in degree, as has been true of their abilities, but in kind. To them questions such as why, how, and whence are less important. To be specific, they are more interested in people as friends, figureheads or freaks than as organisms motivated by needs and drives and acted upon by forces in a human and material environment; a mountain is something to admire, to picnic on or to take a picture of rather

than to analyze as a manifestation of ancient geological goings-on. Two types of special aptitude are needed for success in clerical work, the abilities to recognize verbal and numerical symbols with speed and accuracy. In addition, sales people need certain personality traits which enable them to make effective contact with customers.

The same tests that are used with the prospective academic pupils can be used with those who are considering commercial courses. The two special clerical aptitudes can be measured by well-proved clerical aptitude tests, even at the fourteen-year-old level. Personality cannot so well be measured; tests and inventories are available, some of which have their uses, but it will be some time yet before they are worth the cost for purposes such as those now being considered.

For the trade courses, as for the pre-professional and commercial, a minimum of abstract mental ability is required, but if we may judge by the evidence available from trade schools and from industrial research, the minimum for most trades is lower than for the other two groups. Apparently an I. Q. exceeding 85 or 90 is generally sufficient to enable one to master the arithmetic and other school subjects needed in most skilled trades, although some rate considerably higher than most routine office jobs. Given this, certain special aptitudes assume primary importance, the specific aptitudes and the amounts of those needed varying somewhat from trade to trade. More than average manual dexterity is not needed in most skilled trades, but for those in which it is required, it can be tested. What appears to be most important both in learning a trade and in practicing it is mechanical aptitude or insight, a special ability which is independent of scholastic aptitude. It can be measured fairly well by means of several paper and pencil tests and by performance tests. Of equal importance (and probably underlying the former) is ability to visualize spatial relations; that is, to judge the relationships of shapes and sizes in work such as machine shop and drafting. This ability can be measured by good group and individual

tests Finally there is interest, which must be considered in this as in other fields. The interests of persons in trade schools and in the skilled trades tend to resemble those of people in technical schools and in scientific occupations, but on a lower mental level. They like the concrete and the practical; they prefer to work with objects which they can manipulate and transform rather than with abstract problems, with records, or with people. These interests, and the aptitudes mentioned above, can be measured fairly satisfactorily in early adolescence

We will in time pay more attention to aptitudes needed for education for the semi-skilled and unskilled occupations. Again, minimum and maximum levels of mental ability will need to be taken into account, and these will be lower than for the other types of curricula Achievement in the tool subjects of the school will have less vocational importance. Mechanical aptitude will not play a prominent part. Manual dexterity and ability to visualize spatial relations will vary considerably from one kind of job to another. Physical strength and stamina will play more part in unskilled work, less in semi-skilled. The interests of people who enter these occupations, if we may judge by the not too adequate data now available, are not clearly differentiated. Apparently they have little in the way of special educational or occupational interests. We must study them more intensively in order to find out what does really challenge and appeal to them if we are to devise satisfactory curricula for these groups As we learn more about them we will develop more adequate tests for working with them, especially in the field of interests. The other characteristics can be measured reasonably well at present.

A very important objection is not infrequently raised at this point. Assuming that we use these tests and obtain such information about our pupils, how are we going to get them to act upon it? Are we going to tell them what they can and what they cannot do? Are we going still further and tell them what they may and may not do? Is such action in line with

the democratic philosophy which is one of our basic assumptions?

The answer lies in pointing out that tests can be used for either of two purposes: guidance or selection. The necessity for questions such as the above arises from a confusion of the two, and a consequent misunderstanding of the former. Guidance, or counseling, consists of helping a person to gain insight, to develop self-understanding. Selection involves choosing those who have the desired characteristics and offering them the opportunity in question. In a democratic society we must do both, but the processes must not be confused.

The vocational and educational counselor has, as his function, helping youth to obtain experiences which will give them insight into their abilities and interests. Taking aptitude tests is one such experience. The counselor is concerned also with helping him to evaluate these experiences. Discussing the test results and their significance, as shown in the experiences of others who have made similar scores, is the way in which the results of the test experience are evaluated. The counselor does not say that you can, or you cannot, do thus and so; he shares with the youth information to the effect that such and such a percentage of other youth who made scores comparable to his did or did not do thus and so. The significance of these experience tables must then be discussed, and the youth must make his decision on the basis of the insight thus gained.

The school using aptitude tests for selection is faced with another problem. It, too, has experience tables, to use the insurance term, or norms, to use the educational. Having tested an applicant, it says to him: "You have characteristics which suggest that you will be successful in this line of training and work: we will admit you as a student"; or it says: "Experience shows that most students with your characteristics do not complete our course, so we do not feel justified in investing time and money in giving you this training." Thus, society protects itself and its resources, and experience teaches the individual to make a wiser choice.

APTITUDE TESTING IN PUBLIC SCHOOLS

Such uses of tests imply the existence of two basic conditions: tests which are thoroughly standardized, and test users who know their tools. To administer and to score most tests is relatively easy. To interpret them wisely requires great skill, considerably specialized knowledge, and profound wisdom ripened by experience.

A few brief words of summary may be helpful in closing. The place of aptitude testing in the public schools is the place at which choices need to be made. It is the place at which objective data are needed to provide a basis for those choices. And it is the place at which a trained, skillful, and wise counselor is available to assist in evaluating the data on which those choices should be based.

EFFECT OF ENGINEER SCHOOL TRAINING ON THE SURFACE DEVELOPMENT TEST

RUTH D. CHURCHILL, JEANNE M. CURTIS, CLYDE H. COOMBS,
AND THOMAS W. HARRELL, 1st Lt., A.G.D.

Personnel Procedures Section, The Adjutant General's Office

FAUBION, CLEVELAND, AND HARRELL¹ report that six weeks of "intensive training in mechanical courses does not significantly increase mechanical aptitude test scores, even where the test is very similar to the activities carried out in the training. This is strikingly true of the *Surface Development Test*, in which the items resemble mechanical drafting and blueprint reading work."

An analysis of the effects of nine weeks' training at an Enlisted Men's Engineer School gives contradictory results for the *Surface Development Test*. In this case, there are significant increases in scores on the second administration of the *Surface Development Test* after nine weeks' training.

The content of the *Surface Development Test* used in this study is similar to that of the one used in the previous study; it involves matching drawings in two dimensions and in three-dimensional perspective.

Since the same form of the test was used both at the beginning and at the end of the training, the increases in scores may be attributed to two factors: practice effect and the actual training received in the course. The tests were not given to a control group which received no training, but it is possible to compare the increases in scores of two different classes at the school. Since the training received in the Drafting Class is closely associated with the problems of the *Surface Develop-*

¹R. W. Faubion, E. A. Cleveland, and T. W. Harrell, "The Influence of Training on Mechanical Aptitude Test Scores" *Educational and Psychological Measurement*, II (1942), 91-94

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

ment Test, this class can be used as the experimental group. The instruction in the Water Purification Class covers the principles and applications of electricity and automotive mechanics as well as water purification. Presumably this material is little related to the abilities involved in the *Surface Development Test*, so that this class can be used as a control group. Table 1 compares these two classes with respect to their increases in scores after nine weeks' training.

TABLE 1
MEAN SURFACE DEVELOPMENT TEST SCORES OF THE TWO CLASSES BEFORE
AND AFTER NINE WEEKS' TRAINING

	No.	Mean ₁	Mean ₂	D	σD	$\frac{D}{\sigma D}$
Drafting	66	71.89	93.12	21.23	1.66	12.79
Water Purification	60	52.48	64.18	11.70	1.45	8.07
Drafting vs. Water Purification..		19.41	28.94	9.53	2.20	4.32

The gain made by the Drafting Class on the *Surface Development Test* is significantly greater than that made by the Water Purification Class. Both classes also took tests of mechanical information and comprehension before and after the nine weeks' training. The content of these tests is not related to drafting, although it may be to water purification. The two classes made small but significant gains on both tests; the Water Purification Class gained more than the Drafting Class, significantly so on the mechanical information test. It may be inferred, therefore, that, on the *Surface Development Test*, the greater gain of the Drafting Class as compared with the Water Purification Class is a result of the content of the drafting course.

The most probable explanation of the contradictory results of the two studies lies in the difference in the amount and intensity of the training which each group received. For the airplane mechanics, mechanical drafting and blueprint reading was only one out of five courses; over a period of six weeks, they received 40 hours' instruction in this subject. The Drafting Class at the Engineer School, however, studied nothing but drafting and had almost 400 hours' training in the various phases of that subject.

AN AID TO STUDENT COUNSELORS

RALPH F. BERDIE
University of Minnesota

MUCH TIME is spent in the counseling interview establishing rapport between the interviewer and the student and diagnosing problems of varying complexity. Only after the counselor has obtained clues to and adequately diagnosed the problems of the student may actual therapeutic work proceed. In searching for these clues the counselor often spends a great deal of time asking questions and persuading students to talk about their activities and past experiences. Poor achievement or general dissatisfaction on the part of the student may suggest to the counselor the existence of a problem, but he must then determine if the student is worrying specifically about his health, his inability to get along with his father or his meager social life. When this has been done, the student can then be helped to do something about his problem.

The student who comes to the counselor usually has a complaint. He comes because he is having difficulty in choosing a vocation or is failing his chemistry or is running out of money. These expressed problems demand the attention of the counselor and may provide a starting point for his interview. Most often these complaints are only symptomatic of other problems or else are generalized expressions of several other problems. A student claiming vocational indecision may actually be suffering from lack of information regarding his own abilities and characteristics, a paucity of vocational information and paternal pressure urging him toward a distaste-

ful occupation. A student having trouble with his school work may actually be suffering from poor study habits, inadequate reading skills, and too much outside work. After recognizing a general problem the counselor must make a more specific diagnosis and then initiate treatment.

Many students approach the counselor with a particular orientation dependent upon prevailing stereotypes associated with the counseling program. They come for vocational or educational advice without even considering that they may be able to receive help with some of their other problems. A student may come to the counselor for assistance with his study methods and never think that he might possibly learn how to handle an unpleasant family situation nor realize he should try to do anything about it. He has thought of the counselor as serving only one of the several purposes actually served by that counselor.

To assist the counselor in his diagnosis and to suggest to the student the various functions of the counselor, a problem check list has been developed at the Testing Bureau at the University of Minnesota and used successfully for over one year. The check list consists of thirty-three statements of various problems encountered frequently in student counseling. These problems were obtained from books on counseling (3), (4), and from a survey of case histories of students. The purpose of the list is to facilitate the interview processes and to assist the counselor in determining what problems the student faces. It provides an opportunity for the counselor to approach problems that are often difficult to bring up in the interview and gives the student an opportunity to consider what he wants to talk about before the actual interview.

A more extensive problem check list has been published by Moody (1). His longer list may prove more useful in counseling situations which do not provide a great deal of other information about students. Where much information is obtained through interviews, tests, and questionnaires, a shorter check list of problems is more economical and perhaps

more useful in the interview Wrenn (5) has published a check list to help the counselor in diagnosing and treating problems centering around study habits. Symonds has also done extensive work involving a check list of problems of adolescents and others (2).

The problem check list was included in the individual record form used at the Testing Bureau and given to the students before the counseling interview. Directions to the student were as follows:

Everyone faces problems throughout his life. Some of these problems cannot be solved without help. Many times they are very easily solved. At other times they are solved only after much effort. Below are a list of problems with which young people are often concerned. After those problems you have *not* been able to solve adequately, place a check (✓). After those problems which you would like to discuss with a counselor, place a double check (✓✓). These will help us to be of greater assistance to you.

The responses of 208 men students and 119 women students were tabulated to determine the number of students checking each problem. The number and percentage of men and women checking and double-checking each of the items are presented in Table 1.

Over one-half of the men and women coming to the Testing Bureau desired to discuss what they were best able to do. Slightly less than one-half wanted to discuss what they would like to do. Students coming for counseling express great concern with their abilities and interests, as well they might. The two other items students most wished to discuss were their study habits and the training requirements for their chosen occupations. Students and faculty members have tended to place great emphasis upon the educational and vocational services offered by the Testing Bureau, and problems in these areas are the ones which students are most ready to bring to the counselors.

Comparison of the numbers of students checking and dou-

TABLE 1
NUMBER AND PER CENT OF 208 MEN AND OF 119 WOMEN WHO PLACED SINGLE AND
DOUBLE CHECKS OPPOSITE EACH OF THE PROBLEMS

	Men				Women			
	Single Check		Double Check		Single Check		Double Check	
	No.	%	No.	%	No.	%	No.	%
1. I usually feel inferior to my associates	30	14	10	5	23	20	5	5
2. I have been unable to determine how much time I should study	30	15	31	15	9	8	14	12
3. I have too few social contacts	31	15	5	2	12	10	6	5
4. I have difficulty in making friends	15	7	1	0.5	6	5	4	3
5. I do not know how to obtain the money I need	14	7	11	5	5	4	4	3
6. I have been unable to determine what I am best able to do	50	24	106	51	17	14	57	48
7. I do not know how to take good lecture notes	58	28	29	14	20	17	17	14
8. I do not get along well with my parents	9	4	0	0	8	7	0	0
9. I often have difficulty in keeping friends	8	4	1	0.5	1	1	1	1
10. I am unable to determine what I would like to do	30	14	72	35	20	17	42	35
11. I have not obtained parental approval of my vocational plans	9	4	1	0.5	2	2	2	2
12. I do not have enough to talk about in company	36	17	7	3	21	18	3	3
13. I receive inadequate financial help from my family	7	3	2	1	3	3	2	2
14. I do not know how to outline text-book assignments	26	13	12	6	4	3	9	8
15. I am unable to get along with my brothers and/or sisters	3	1	1	0.5	3	3	0	0
16. I have been unable to make a satisfactory religious adjustment	15	7	1	0.5	9	8	1	1
17. I am not interested in my studies	13	6	3	1	1	1	3	3
18. I do not have enough information about job opportunities and duties	24	12	27	13	15	13	10	8
19. I am frequently embarrassed when with others	17	8	1	0.5	10	8	2	2
20. I usually do not enjoy being with members of the opposite sex	10	5	3	1	4	3	1	1
21. I am unable to do my work well because of too many social activities	9	4	2	1	4	3	1	1
22. I usually do not know how to act in company	10	5	0	0	0	0	0	0
23. I usually cannot read fast enough to cover all of my assignments	23	11	10	5	10	8	1	1
24. I usually have difficulty understanding what I read	19	9	4	2	12	10	1	1
25. I do not know what the most appropriate training is for my chosen career	17	8	39	19	7	6	24	20
26. I do not know if an education is worth while	5	2	6	3	2	2	0	0
27. I feel guilty about something I have or have not done	14	7	1	0.5	8	7	3	3
28. I have so much outside work to do that I am neglecting my school work	3	1	3	1	1	1	1	1
29. I have trouble making myself study	51	25	19	9	12	10	9	8
30. I lack self-confidence	35	17	9	4	29	24	10	8
31. I am dissatisfied with my state of health	12	6	2	1	4	3	0	0
32. I do not know how to improve my personal appearance	5	2	0	0	1	1	0	0
33. I do not know how to break certain habits I have	12	6	1	0.5	4	3	1	1

ble-checking each item reveals that many students are aware of problems which they are not eager to discuss with a counselor. Many students feel inferior to their associates but do not express a desire to discuss this with a counselor. Many state that they lack self-confidence. Many consider that they have too few social contacts but would not like to talk to a counselor about this. In view of the many techniques available for the counselor in dealing with the social problems of students at the University, the students appear to be turning away from possible assistance. Relatively more students single-check items related to reading problems than double-check these items. Inspection of these figures shows that students are aware of their educational and vocational problems and are frequently willing to talk about them. Although they often recognize social problems and personal problems they have little desire to discuss these with their counselors.

This reluctance to discuss certain types of problem may be due to the fact that the students think that nothing can be done about these problems and that consequently time would be wasted in discussing them with a counselor. They may consider their personal problems too private to discuss with a relative stranger, but this would hardly explain the few double checks opposite the reading problems. When students come to the counselor, they come with one primary purpose and all other matters may appear irrelevant at that time. Coming to a counselor may follow or accompany some crisis in the life of a student, a failure or a forced change of vocational plans, and this crisis may envelop the entire horizon of the student.

The students included in this study had also filled out the Minnesota Personality Scale. On this test scores are available for morale, social adjustment, family adjustment, emotional adjustment, and economic conservatism. The score on the morale section is related to the individual's emotional acceptance of surrounding social and community situations. Very high scores may indicate naïve optimism, low scores cynicism or lack of hope for the future. Scores on the social ad-

justment section are related to the social maturity, gregariousness, and socialization of the individual. High scores on the family relations section indicate friendly and healthful child-parent relations. Scores on the emotionality section are related to emotional stability. Low scores on this section may often result from hypochondriasis, anxiety states, or over-reactive tendencies. Scores on the economic conservatism section are related to the liberality of the individual's economic attitudes.

On the basis of selected personality test scores, comparisons were made between the means of those students who had checked items and those who had not checked those same items. For example, of the 193 men for whom complete data were available, 11 checked the item, "I do not know if an education is worth while." This item was left unchecked by 182 students. The mean personality scores for groups checking and not checking the items and their critical ratios (difference divided by standard error of the difference) are presented in Table 2. The items have been grouped into functional categories on an inspectional or logical basis. Only those items have been included which appeared relevant to the scores on the Personality Scale.

On each item related to social behavior, the group checking the item had significantly lower scores on the social adjustment test than did the group not checking the item. The men who indicated that they had too many social activities, however, had higher scores on the social adjustment section, as would be expected. Since many of the problems on the check list are very much like some of the items on the test, the observed relationship is not surprising. Students who claim that they do not get along well with their parents obtain significantly lower scores on the family relations section of the personality scale than students who do not check this item. However, men whose parents do not approve of their vocational choice do not obtain significantly lower scores on this section. Students who claim that they have been unable to make a satisfactory religious adjustment make no lower

TABLE 2

COMPARISON OF MEAN SCORES ON PERSONALITY SCALE OF STUDENTS CHECKING AND NOT CHECKING PROBLEMS

M E N				
Section of Personality Scale	Problem	Mean of Students Who Checked Item	Mean of Students Who Did NOT Check Item	Critical Ratio
<i>Morale</i>				
	I do not know if an education is worth while156.55	157.45	20
<i>Social</i>				
	I have too few social contacts196.83	223.66	4.56
	I have difficulty in making friends184.88	221.71	3.72
	I do not have enough to talk about in com- pany192.19	226.01	5.74
	I am frequently embarrassed when with others182.06	222.19	3.70
	I usually do not enjoy being with members of the opposite sex187.54	220.90	2.94
	I am unable to do my work well because of too many social activities239.80	217.50	2.86
	I usually do not know how to act in com- pany177.10	220.92	2.98
	I lack self-confidence197.95	224.07	4.10
<i>Family</i>				
	I do not get along well with my parents.99.66	120.06	2.59
	I have not obtained parental approval of my vocational plans109.78	119.56	1.44
<i>Emotional</i>				
	I have been unable to make a satisfactory religious adjustment134.36	129.77	.94
	I am frequently embarrassed when with others119.41	131.13	2.20
	I usually do not enjoy being with members of the opposite sex.128.23	130.23	.42
	I feel guilty about something I have or have not done118.57	131.00	2.26
	I lack self-confidence121.70	132.29	3.15
	I am dissatisfied with my state of health.116.85	131.06	2.39
W O M E N				
<i>Social</i>				
	I have too few social contacts.182.47	201.19	2.29
	I have difficulty in making friends.162.56	201.90	5.20
	I have not enough to talk about in com- pany177.00	203.96	3.99
	I am frequently embarrassed when with others179.09	200.78	2.16
	I lack self-confidence179.97	207.58	4.76
<i>Family</i>				
	I do not get along well with my parents102.38	139.02	6.76
<i>Emotional</i>				
	I have been unable to make a satisfactory religious adjustment161.29	163.70	.29
	I am frequently embarrassed when with others146.64	165.55	3.10
	I feel guilty about something I have or have not done148.10	165.19	2.34
	I lack self-confidence.158.55	165.89	1.51

scores on the emotional adjustment scale than do other students. Guilt feelings apparently are related to the score on this section, and both men and women who check this item obtain significantly lower scores than those who do not check it. Men who claim dissatisfaction with their state of health obtain lower emotional adjustment scores, as do men who claim to lack self-confidence. Women who indicate a lack of self-confidence, however, do not differ significantly on the basis of this scale from other women.

Of the 27 relationships analyzed, 21 were found to be statistically significant. Students tending to obtain low scores on the various sections of the personality scale will also tend to check related items on a problem list and thus supply the counselor with a clue regarding the source of these low scores. Perhaps the same information could be obtained by going through the items of the test, but as there are 218 items in the test, each with five possible answers, this would require much of the counselor's time. A factor analysis of the items of the test would perhaps identify a few key items which could be used for the same purpose as the problem check list, but until this is done, the counselor may more economically glance at the 33 items on the check list than read through the many items of a personality test.

Added to the statistical evidence concerning usefulness of the problem check list is much evidence obtained from clinical work involving the use of this instrument. The description of a few cases in which it has proved useful will exemplify this and also suggest various techniques that have proved successful in using the list in the interview.

Joseph H. came to the Testing Bureau for assistance in deciding upon a major in the College of Education. He was completing his second year in the university and had been doing slightly better than average work. He had graduated from a small high school, and his social life had been very restricted in the little town from which he came. Among other items, he double-checked that he had too few social contacts. His percentile score on the social adjustment section

of the Personality Scale was 24. After discussing the boy's vocational plans, the counselor said, "I see that you check that you have too few social contacts. What do you think you could do about that?" Joseph started to discuss the facilities available on the campus and soon he and the counselor had a social program planned, and the counselor gave him a letter of introduction to the secretary of the Y. M. C. A. The item checked by the boy in this case gave the counselor an opportunity to approach a problem the existence of which might have been easy to determine but for which treatment might not have been initiated so easily without the item.

George S had checked several items on the problem check list, including the one, "I feel guilty about something I have or have not done." A single check had been placed opposite this item. During the interview the counselor decided that the boy presented a picture of a very unstable individual and that various personal problems might interfere with his progress when he entered college the following fall. The counselor was unable, however, to get the boy to discuss these problems. Finally, he said, "I see you checked here that you feel guilty about something you have done or have not done." After a pause he continued, "Many people feel guilty about things they have done, and usually feeling guilty about it is the only thing that does any harm." He paused again, and George began to speak of the problems that had been worrying him and of his reactions to these problems. In this case, the problem check list provided an opportunity for the counselor to approach a problem which had previously resisted all attempts to approach it.

We have found that when an item is double-checked, the most convenient and profitable thing the counselor can do is to refer directly to the item and give the student an opportunity to elaborate upon his response. When only a single check is placed opposite the item, however, this can seldom be done. The counselor will have to remember that the student did not indicate that he wanted to talk about the subject checked and that he may actually resent any attempt on the

part of the counselor to start such a discussion. When an item has been checked only once, the counselor can often discuss the problem mentioned and give the student an opportunity to ask questions without making the student aware that the item itself is being referred to.

Sarah W. placed a single check opposite the item, "I have not obtained parental approval of my vocational plans." During the interview, after discussing various alternatives, the counselor asked, "What do you think your parents would like you to do?" Sarah then told what her parents' reactions were and also revealed a family problem which had not even been suspected up to that point in the interview. If the counselor had asked, "Why don't your parents approve of your vocational plans?", it is doubtful if Sarah would have given the information she actually gave.

Summary

Statistical analysis of a problem check list and its clinical use have shown that it is a useful instrument in diagnosing students' problems and in approaching these problems in the interview. The items checked offer the counselor an opportunity to select those areas which offer most promise for investigation and to introduce these topics in the counseling interview. The items also assist in orienting the student toward the counselor and in reaching a definition of his problems before the interview.

REFERENCES

- (1) Moody, R. *Problems Check List* Columbus, Ohio: Ohio University Press, 1941.
- (2) Symonds, P. K. "Life Problems and Interests of Adolescents," *School Review*, XLIV (1936), 506-518.
- (3) Williamson, E. G. *How to Counsel Students* New York: McGraw-Hill, 1939.
- (4) Williamson, E. G. and Dailey, J. G. *Student Personnel Work*. New York: McGraw-Hill, 1937.
- (5) Wrenn, C. G. *Study-Habits Inventory*. California: Stanford University Press, 1941.

A COMPARISON OF THE HUMAN BEHAVIOR INVENTORY WITH TWO OTHER PERSON- ALITY INVENTORIES

ABRAHAM SPERLING
City College of New York

PENCIL-AND-PAPER TESTS for diagnosing personality traits have too frequently proved unsatisfactory to the investigators employing them. Statements expressing discontent with the diagnostic results of such tests are found in studies by Watson (1), Mosier (2), Landis (3), Moore and Steele (4), Feder and Mallet (5), Gorham and Brotemarkle (6), Stagner (7), and others too numerous to include here. Accompanying the criticisms, however, constructive suggestions are frequently made for the improvement of such instruments. Among the suggestions offered are the use of multiple answers, the use of weighted scoring, the development of reliability, better definition of terms, and abstention from scoring the same items for more than one trait. Because it is felt that the *Human Behavior Inventory*,¹ devised by Randolph B. Smith, represents an improvement in adjustment scales in accordance with these suggestions, it is the desire of the investigator to bring this instrument to the attention of possible users.

Employed in an experimental study (8) conducted by the investigator, the *Human Behavior Inventory* proved to be a most satisfactory instrument for measuring traits of personality adjustment. It was devised for the purpose of testing personality adjustment of a group of college students. Smith developed the instrument by selecting from previous inventories the items found most diagnostic, modifying them in an effort

¹This inventory is reproduced in a monograph by Smith (9)

to make them capable of measuring status as well as change, and adding new items where necessary. In his original study, Smith (9) employed the test to measure the personality status of college students at the beginning and end of a school year in which the individuals had been subjected to a course in mental hygiene.

Description of the Test

The inventory was developed to yield a total score which might serve as a measure of general personality adjustment, together with separate subscores on six individual sections (1—work efficiency, 2—superiority-inferiority or degree of self-confidence, 3—social acceptability and adjustment, 4—emotional stability with reference to neurotic symptoms, ease of adjustment to new experiences, and general sex adjustments, 5—objectivity toward behavior of others, and 6—family attitudes and relationships) which may be regarded as major characteristics of mental health and emotional maturity. The scale contains 102 questions, answers to which are based on a five-degree multiple choice. Each item is given a score ranging from 0 to 4, depending upon the degree of the answer. The estimated reliability of the total test score by odd-even correlation for 1125 cases was $.89 \pm .01$ and by test-retest correlation for 465 cases after six months was $.81 \pm .01$.

Procedure

This investigation was undertaken to compare the *Human Behavior Inventory* with previously constructed scales. Several of the instructors in elementary psychology at the College of the City of New York administered to their classes the *Human Behavior Inventory*, the *Bell Adjustment Inventory* (10), and the short form of the *Thurstone Personality Schedule* as revised by R. R. Willoughby, which is known as the *Clark-Thurstone Inventory* (11). Statistical data concerning these scales are described in the bibliographical references noted.

To each class in elementary psychology both the *Human Behavior Inventory* and the *Clark-Thurstone Inventory* were

HUMAN BEHAVIOR INVENTORY

given during the same period. The *Bell Inventory* was given during the subsequent class meeting. One hundred seven complete sets of inventories were made available to the investigator.

The Data

The intercorrelations of the three scales are presented in Table 1, while other statistics concerning each inventory are given in Table 2.

TABLE 1

INTERCORRELATIONS AMONG HUMAN BEHAVIOR INVENTORY, BELL INVENTORY, AND CLARK-THURSTONE INVENTORY

Inventories	Coefficient of Correlation	No. of Items in Respective Scales	No. of Identical Items Between Scales
Human Behavior Inventory and Bell Inventory	.736 \pm .030	102 140	12
Clark-Thurstone and Human Behavior Inventory	.748 \pm .029	25 102	7
Bell Inventory and Clark-Thurstone	.785 \pm .026	140 25	18

TABLE 2

SCORES ON HUMAN BEHAVIOR INVENTORY, BELL INVENTORY, AND CLARK-THURSTONE INVENTORY^a

Scales	Range	Mean	S. D.	Coefficient of Reliability
Human Behavior Inventory	36-202	120 (123)	38.55 (39.23)	.918 (.89)
Bell Inventory	1-77	33.2 (32)	16.50	(.93)
Clark-Thurstone Inventory	2-67	26 (29)	15.07 (13.70)	(.91)
Age	17-24.6	19.5		N = 107

^aFigures in parentheses are from the original studies by the authors.

The fact that the *Human Behavior Inventory* correlates rather highly with the two older scales should not give the impression that it is a duplication of the exact content of the scales with which it was compared. However, the high correlations probably indicate that it tends to measure the same factors, namely, those of personality adjustment. To check whether the high correlations were due to mere identity of the items in the scales, an analysis of the three questionnaires was made.

The analysis (summarized in Table 1) showed that there were eighteen identical items in the Bell and Clark-Thurstone, seven in the *Human Behavior Inventory* and the Clark-Thurstone, and twelve in the *Human Behavior Inventory* and the Bell. While it is thus seen that exact identity of items alone does not provide a reasonable explanation of the fairly high correlations between the scales, it is of course recognized that similarity of items not identical may also be a factor.

The intercorrelations among the sections of the scales have been reproduced for two reasons: first, to offer a record of the data for the benefit of others who may wish to make comparisons and, second, to demonstrate the similarity of results obtained in this study and in the original studies by the respective authors. To illustrate, Tables 3 and 4 show the closeness of the mean scores and intercorrelations obtained from the 107 subjects of this study to those of the 1145 of Smith's study and the 258 of Bell's study. The rather high interrelationships among the parts of the *Human Behavior Inventory* may be an indication that the sub-scores do not represent separate psychological factors. However, it is possible that they are indicative of truly close relationships among the several personality traits measured by the subsections. Further exploration of these possibilities may well be the subject of a subsequent investigation.

It may be pertinent to mention at this point that in the opinion of the investigator the importance of establishing rapport between experimenter and subject for best results from pencil-and-paper tests of personality cannot be overempha-

HUMAN BEHAVIOR INVENTORY

TABLE 3

CORRELATIONS OF PARTS OF BELL INVENTORY WITH EACH OTHER
AND WITH TOTAL*

Parts	Health	Social	Emotional	Total
Home	.39 (.43)	.21 (.04)	.54 (.38)	.757
Health		.18 (.24)	.45 (.53)	.629
Social			.44 (.47)	.655
Emotional				.832

*Figures in parentheses are from the original study by the author.

TABLE 4

CORRELATIONS OF PARTS OF HUMAN BEHAVIOR INVENTORY WITH
EACH OTHER AND WITH TOTAL*

Parts**	Sup Inf	Soc. Acc.	Emot Stab	Obj	Fam. Rel	Total	Total Less This Sec.	M	S. D	No. of Items in Sec
Work	.60 (.59)	.47 (.52)	.56 (.56)	.42 (.40)	.38 (.44)	.58 (.68)	.508	12.87 (12.94)	4.81 (4.90)	
Eff		.68 (.66)	.73 (.65)	.59 (.46)	.43 (.46)	.78 (.76)	.772	14.12 (15.57)	5.86 (5.96)	9
Sup			.76 (.72)	.56 (.47)	.53 (.54)	.80 (.80)	.797	16.86 (17.75)	6.66 (7.05)	11
Inf				.66 (.63)	.64 (.60)	.92 (.88)	.908	28.59 (30.22)	11.20 (11.55)	13
Soc					.45 (.55)	.77 (.74)	.615	18.58 (18.08)	7.43 (6.93)	29
Acc.						.78 (.85)	.607	29.43 (28.50)	12.26 (11.85)	14
Emot										26
Stab.										
Obj										
Fam										
Rel.										

*Figures in parentheses are from the original study by R. B. Smith.

**The abbreviations of the subsection names refer to Work, Efficiency, Superiority-Inferiority, Social Acceptability, Emotional Stability, Objectivity, and Family Relationships.

sized In this study, extreme care was taken in the matter of rapport In the original instructions each student was asked to volunteer his efforts in a research study that would have no bearing on his grades or standing at the college. Each indi-

vidual was told that his replies would be treated entirely confidentially and anonymously unless he desired otherwise. He was asked to be sincere and objective and to inform the investigator if he felt his rapport was not valid. As an added incentive for an honest expression of their own characteristics as they know them, the students were told that they would be given the results of their tests in such a manner that they could compare their scores with the average of others taking part in the study if they so desired. A majority of the students were known to the investigator in a rather friendly student-teacher relationship. It is the opinion of this investigator that disappointing results from the use of personality questionnaires are frequently due to a lack of rapport between experimenter and subjects.

Summary and Conclusion

The coefficient of correlation between the *Human Behavior Inventory* and the *Bell Inventory* was .736, that between the *Human Behavior Inventory* and the *Clark-Thurstone Inventory* .748, and that between the *Bell* and the *Clark-Thurstone* .785. An analysis of the three measures showed more similar items between the *Bell* and the *Clark-Thurstone* scales than between the *Human Behavior Inventory* and either of these measures.

In view of the similar positive coefficients of intercorrelation among the three scales, it may be concluded that the *Human Behavior Inventory* is probably as satisfactory for use as a diagnostic measure of personality adjustment as either of the other two measures with which it was compared. Moreover, the scale embodies several desirable features such as the use of multiple answers, weighted scoring, high reliability, clear definition of terms, and abstention from scoring the same items for more than one trait. Since these aspects are among the suggestions made by authorities for the improvement of personality scales, they lend support to the acceptance of the *Human Behavior Inventory* as an instrument for measuring traits of personality adjustment.

HUMAN BEHAVIOR INVENTORY

REFERENCES

1. Watson, G. "Personality and Character Measurement," *Review of Educational Research*, VIII (1938), 269-291.
2. Mosier, C. I. "On the Validity of Neurotic Questionnaires," *Journal of Social Psychology*, IX (1938), 3-16.
3. Landis, C. "Empirical Evaluation of Three Personality Adjustment Inventories," *Journal of Educational Psychology*, XXVI (1935), 321-330.
4. Moore, H. and Steele, I. "Personality Tests," *Journal of Abnormal and Social Psychology*, XXIX (1934), 45-52.
5. Feder, D. and Mallet, D. "Validity of Certain Measures of Personality Adjustment," *Journal of American Association of College Registrars*, XIII No. 1 (1937), 5-15.
6. Gotham, D. R. and Brotmarkle, R. "Challenging Three Standardized Emotionality Tests for Validity and Employability," *Journal of Applied Psychology*, XIII (1929), 554-588.
7. Stagner, R. "The Intercorrelation of Some Standardized Personality Tests," *Journal of Applied Psychology*, XVI (1932), 453-464.
8. Sperling, A. *The Relationship between Personality Adjustment and Achievement in Physical Education Activities*, Doctoral dissertation, 1941. On file in the library of New York University, New York.
9. Smith, R. B. *Growth in Personality Adjustment Through Mental Hygiene*, Albany, New York: University of the State of New York, State Education Department, 1936.
10. Bell, H. M. *Manual for the Adjustment Inventory*, Stanford University, California: Stanford University Press, 1934.
11. Willoughby, R. R. "Some Properties of the Thurstone Personality Schedule and a Suggested Revision," *Journal of Social Psychology*, III (1932), 401-424.

INTRA-INDIVIDUAL DIFFERENCES VERSUS INTER-INDIVIDUAL DIFFERENCES IN MOTOR SKILLS¹

WILLIAM A. OWENS, JR.
Iowa State College

STUDIES OF VARIATION within and between individuals have tended to be restricted to one sort of intra-individual variation, trait differences. They have also tended, in treating of the relative magnitudes of individual differences and trait differences, to display adherence to one of two modes of attack. Either they have dealt with the inter-correlations of certain traits or functions, or they have shown a comparison of a trait standard deviation with a standard deviation representative of individual differences.

The present paper is an attempt to evaluate trait differences and certain other intra-individual factors, and to relate them in magnitude to individual differences.

The writer feels that the statistical technique which was employed in the present investigation was superior to either of the two which are conventionally used for the reasons which follow. First, neither the method of inter-correlation nor the method of comparing standard deviations will allow of the treatment of more than one intra-individual factor. Second, even in the comparison of individual and trait differences, the

¹This article is a condensation of the writer's doctoral dissertation of the same title, a copy of which is on file at the library of the University of Minnesota.

The writer wishes to acknowledge the invaluable criticisms and suggestions of his advisors, Professor D. G. Paterson and Dr. P. O. Johnson. He also wishes to recognize the assistance of Dr. Brent Baxter and of Mr. Paul G. Homeyer.

The actual experimental work was done in the psychological laboratories at the University of Minnesota with the cooperation of Dr. M. A. Tinker, and was made possible through a research grant by the Graduate School of that institution.

magnitude of the correlation coefficient is conditioned by at least two variables besides the true magnitude of the relationship. These are unreliability of measurement and trait variability.² Third, product-moment correlation, or its equivalent, deals with absolute *deviate* ranks; no account of variation within these rank positions is taken so long as they do not actually shift.³ Fourth, standard deviations are increased by unreliabilities of measurement. If a systematic error were present, the error in measuring trait differences would not be equal to the error in measuring individual differences. Even if this were not the case, a constant error factor would be a relatively larger component of trait differences—if it were the smaller—than of individual differences. A statement of proportionality could, thus, only be accurate if individual and trait differences were of the same magnitude.

The present experiment, which was designed for application of the analysis of variance, was planned with a view to taking account of these several objections to other techniques. In accordance with this purpose, certain facts are worth noting. First, account is taken of more than one intra-individual factor. Second, various sorts of unreliability are incorporated in the estimate of error with at least two relevant consequences: the estimates of the magnitudes of the main factors are correspondingly more accurate, and direct tests of

²Since terminology in this field has not been entirely uniform, the writer includes his own definitions of the terms he employs.

Inter-individual differences = differences in relative proficiency from individual to individual. *Intra-individual differences* = (1) trait differences, (2) repetitive variations, (3) trait variability, et al.

Trait differences = differences in relative proficiency from function to function within the individual.

Repetitive variations = changes in the individual's proficiency from day to day in the average of all functions measured. The systematic portion of this shift might be designated as learning, and the random portion attributed to shifts in the subject's efficiency.

Trait variability = the term used by Paulsen to denote the fluctuation of a given function within an individual, temporally. See Paulsen, G. B. "A Coefficient of Trait Variability," *Psychological Bulletin*, XXVIII (1935), 218-19.

³Harris has taken account of such a contention in developing his method of relative correlation. The procedure is to correlate a first variable with the deviation of a second from its most probable value. Harris, J. A. "The Correlation Between a Variable and the Deviation of a Dependent Variable from its Most Probable Value," *Biometrika*, VI (1908), 438-443.

INTRA-INDIVIDUAL DIFFERENCES

the significances of these factors are made possible. Third, the analysis is so based as to minimize the errors normally incurred in the ranking of data, while the statistical technique employed takes account of the total variation—none of it goes exempt from analysis. With this brief preface, the present experiment may be outlined.

The Problem—To obtain an estimate of the relative magnitudes of individual differences and of several intra-individual factors on some tests of motor skills.

The Method—Table 1 provides an abbreviated illustration of the technique employed to determine the per cent of the total variation in score contributed by individual differences.

TABLE 1
A SAMPLE ANALYSIS

Individuals (15)	Administrations—Block Packing				VIII
	II	III	IV	.	
A	425	448	425		
B	502	469	530		
C	514	541	648		
D	384	403	392		Norm
E	345	436	463		M = 500
F	512	467	577		$\sigma = 100$
G	511	611	623		
H	241	372	428		
I	522	573	548		
J	453	498	567		
K	317	380	302		
L	461	547	517		
M	394	478	496		
N	534	567	631		
O	381	441	491		

Correction Term = $(21365)2/45 = 10,143,627.22$

Total Variation = 356,569.78

Individual Differences = 272,389.11

Repetitive Variations = 44,667.51

Error (Interaction) = 39,513.16

Factor	Degrees of Freedom	Sum of Squares	Mean Square	"F"	P	%
T.V.	44	356,569.78				100
I.D.	14	272,389.11	19,456.37	13.79	< .01	68
R.V.	2	44,667.51	22,333.76	15.83	< .01	16
Err.	28	39,513.16	1,411.18			16

The analysis of variance was employed with two criteria of classification — individuals and administrations.⁴ The isolates from the total variation (T.V.) were individual differences (I.D.),⁵ repetitive variations (R.V.), and an estimate of error (Err.) Especially to be noted is the fact that the per cent column furnishes an estimate of the magnitude of the contribution of individual differences to the total variation. An analysis of this sort was run for each test of the present experiment.

The analysis of intra-individual factors took a similar form. Table 2 illustrates the method and the character of the second series of analyses. The isolates from the total variation (T.V.) were trait differences (T.D.), repetitive variations (R.V.), and error (Err.). An analysis of this sort was made for each subject in the experimental group. The intention was ultimately to compare the per cent contribution of individual differences to the total variation in the first series of analyses with the respective per cent contributions of trait differences and repetitive variations to the total variation in the second series of analyses.

Seven tests of motor skills were employed in this investigation. They were the block packing, steadiness, speed of movement, slow movement, stick balancing, tapping, and card sorting tests reported in the Minnesota Mechanical Ability Study. Each one of 15 subjects was given eight administrations of each of the seven tests — 56 testings per subject.⁶ The tests were administered in systematically varied order on a schedule calling for about eight hours of each person's time. Subjects were junior high school boys matched for age, intelligence, and race, both with each other and with a norm group of 216 individuals.⁷ These boys were paid for their time, and

⁴C. H. Gouldeen, *Methods of Statistical Analysis* (New York: John Wiley and Sons, 1939), pp. 114-141; especially p. 127.

⁵From this type of analysis only the individual differences factor figures in later comparisons.

⁶Results from all first administrations were, as usual, discarded as unreliable.

⁷D. G. Paterson, R. M. Elliott, et al. *Minnesota Mechanical Ability Tests* (Minneapolis: University of Minnesota Press, 1930), p. 586.

INTRA-INDIVIDUAL DIFFERENCES

TABLE 2

A SAMPLE ANALYSIS

Tests (6)	Administrations — Subject E				. . .	VIII
	II	III	IV			
Block Packing	345	436	463			
Steadiness	370	370	306			Norm
Slow Movement	512	618	583			M = 500
Speed of Movement	652	712	649			$\sigma = 100$
Tapping	557	601	619			
Stick Balancing	520	525	538			
Correction Term	= (9376)2/18 = 4,883,854.22					
Total Variation	= 234,357.78					
Trait Differences	= 213,415.11					
Repetitive Variation	= 8,069.78					
Error (Interaction)	= 12,872.89					

Factor	Degrees of Freedom	Sum of Squares	Mean Square	"F"	P	%
T.V.	17	234,357.78				100
T.D.	5	213,415.11	42,683.02	33.16	<.01	89
R.V.	2	8,069.78	4,034.89	3.13	>.05*	3
Err.	10	12,872.89	1,287.29			8

*See a later reference on combining independent probabilities.

several prizes were awarded at the completion of the testing. Motivation appeared to be excellent.

Two methodological issues now demand attention. First, in order to evaluate trait differences — differences in the subjects' relative proficiency from test to test — the various tests themselves had to be equated. This was accomplished in the following fashion. The seven norm distributions of 216 cases each were checked for normality, and the five which departed from the criterion were normalized. The pertinent data are included in Tables 3 and 4. These distributions were then assigned comparable scores after the method described by Hull, McCall, et al.⁸ Specifically, each distribution was assigned a mean of 500 and a standard deviation of 100. The scores of the subjects in the present experimental group were converted to this form and evaluated in terms of these equated

⁸C L Hull, "The Conversion of Test Scores into Series Which Shall Have Any Assigned Mean and Degree of Dispersion," *Journal of Applied Psychology*, VI (1933), 298-300.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 3

NORMS

Test	M	σ	Trans- formation	Value	N
<i>Block Packing*</i>	2.739	.064	logarithm	low score is good	217
T.S.V.**	549	83			
O.S.V.***	571	87			
<i>Slow Movement*</i>	1.518	.349	logarithm	low score is good	217
T.S.V.	22	26			
O.S.V.	41	34			
<i>Steadiness*</i>	2.382	.500	square root	high score is good	217
T.S.V.	5.67	2.38			
O.S.V.	6.18	2.48			
<i>Tapping*</i>	20.514	.898	square root	high score is good	217
T.S.V.	421	34			
O.S.V.	426	37			
<i>Stick Balancing*</i>	1.089	.205	square root of logarithm	high score is good	216
T.S.V.	15	21			
O.S.V.	35	72			
<i>Speed of Movement</i>	150.06	30.138	none	high score is good	217
T.S.V.	150.06	"			
O.S.V.	150.06	"			

* = value in transformed distribution.

**T.S.V. = transformed score value.

***O.S.V. = original score value.

norm distributions. Any tendency to minimize the magnitude of trait differences may thus be viewed as a function of the sampling error of a mean of 216 cases.⁹

The second methodological consideration was the time-honored one of securing a zero point and equal units of measurement on the scale for the evaluation of individual differ-

⁹A trial analysis of the scores on the speed of movement test was made using both the original and the transformed measures. The relative magnitudes of the respective factors were identical to two decimal places by the two methods. Apparently, none of the information latent in the data is lost through the transformation.

INTRA-INDIVIDUAL DIFFERENCES

TABLE 4
TESTS OF NORMALITY

Tests	N	G_1	G_2	σG_1	σG_2	P
Block Packing	217	0.315	0.529	0.1655	0.3286	>.01
Slow Movement	217	0.761	0.845	"	"	"
Speed of Movement . .	217	0.312	0.049	"	"	"
Tapping	217	0.169	0.356	"	"	"
Steadiness	217	0.362	0.144	"	"	"
Stick Balancing	216	0.040	0.145	"	"	"
Card Sorting	219	0.253	0.017			later omitted

G_1 and G_2 are calculated from R. A. Fisher's K statistics. A complete description of the method is to be found in Goulden, C. H. *Methods of Statistical Analysis* (New York: John Wiley & Sons, 1939), pp. 27-31.

ences. This would, of course, be necessary in order to justify the ultimate pooling of the results. Anastasi¹⁰ has pointed out that standard scores from scaled, or normal, distributions tend to yield such equal units of measurement. Also, the analysis of variance deals only with deviates or differences, and not with absolute magnitudes. These two considerations seem to point to the adequacy of the present data and technique for the purpose in view.

It would have been ideal to establish the normality of the distribution of trait differences within each individual in similar fashion. However, the number of traits measured was so small that this constituted a practical impossibility. Hull¹¹ has stated that the distribution of trait differences appears to be a normal one. The data of the present study would affirm this opinion, although no conclusions may be based on the inspectional method employed. In any case, the error, if any, introduced by assuming the normality of the distribution of trait differences would be very slight.

On the assumption that a satisfactory estimate of the relative magnitudes of individual and trait differences might be obtained, an attempt was made to isolate certain other sources

¹⁰A. Anastasi, "Practice and Variability," *Psychological Monographs*, XIV (1933-34), No. 5.

¹¹C. L. Hull, "Variability in Amount of Different Traits Possessed by the Individual," *Journal of Educational Psychology*, XVIII (1927), 97-106.

of intra-individual variation. In the second type of analysis, illustrated in Table 2, the repetitive variations factor (R.V.) is seen to be composed of differences between the means of administrations. A direct estimate of the magnitude of this factor was obtained, as in the case of trait differences, by determining its mean per cent contribution to the total variation from each separate analysis of the second type.¹² However, the differences between the means of the administrations may be viewed as being attributable to two distinct sources: first, learning; and second, random fluctuations in the individual's efficiency from day to day in all functions. An attempt was made to differentiate between the two.

Briefly, it was assumed that learning would be at a maximum on administrations 2-4, and at a minimum and negligible level on administrations 6-8. The evidence for this assumption follows. (1) If the proposed dichotomy in administrations (2-4 vs. 6-8) is made, the repetitive variations factor is significant in the second summary analysis of administrations 2-4, and is insignificant in the second summary analysis of administrations 6-8. Table 5 gives the relevant data, and the probability (P) column illustrates the point. (2) Also to be noted in Table 5 is the fact that the repetitive variations mean square is only slightly larger than the error mean square for administrations six through eight. (3) The establishment of a common unit for the estimation of improvement makes it apparent that most learning is confined to administrations 2-5. The " t " test is generalized in Fisher's concept of fiducial probability to yield an expression as to the magnitude which any difference must attain to be "significant" at any given level of probability.¹³ Specifically, in the present instance, the fiducial limits at the 10 per cent

¹²One analysis for each subject in the experimental group; each one in the form illustrated in Table 2. It makes no essential difference in the results whether a per cent is obtained in each analysis and the mean of the series obtained, or whether the sums of squares and degrees of freedom are totaled and one per cent computed from these "summary statistics." The latter method is probably preferable for purposes of estimation.

¹³R. A. Fisher, *Statistical Methods for Research Workers* (London: Oliver & Boyd, 1937)

INTRA-INDIVIDUAL DIFFERENCES

TABLE 5

REPETITIVE VARIATIONS LEARNING VS. RANDOM FLUCTUATIONS

Factor	Degrees of Freedom	Sum of Squares	Mean Square	"F"	P	%
Administrations 2-4						
T.V.	255	2,828,723.05				100
T.D.	75	2,441,890.33	32,558.54	18.68	<.01	83
R.V.	30	125,390.68	4,179.68	2.40	<.01	3
Err.	150	261,442.04	1,742.95			14
Administrations 6-8						
T.V.	255	2,861,646.86				100
T.D.	75	2,552,178.81	33,629.05	18.22	<.01	85
R.V.	30	62,656.82	2,088.56	1.13	>.05	0.3
Err.	150	276,811.23	1,845.41			15

*These are summary statistics derived by totaling the sums of squares and degrees of freedom of the separate analyses.

level were used as a qualitative, common unit for the measurement of improvement from administration 2 through administration 8. Table 6 contains a summary on this point.

It should be noted that "improvement" in each individual case is defined in terms of the amount of variation which may be viewed as "random." In accordance with the previously stated hypothesis, it will be observed that there is only *one* exception to the rule that learning, if present, tends to be confined to the first five administrations.¹⁴ In view of the evidence presented, it was assumed that the difference between the initial (2-4) and final (6-8) magnitudes of the repetitive variations factor might furnish an estimate of the relative importance of learning.

Finally, it can be shown that the estimate of error, or interaction, in the second series of analyses may have as many as three experimental components. These are: (1) unreliabilities of measurement, presumably inherent in the test; (2) trait variability, presumably inherent in the individual; and (3) differential rates of improvement within the individual

¹⁴It seemed best to omit administration 5 because it appeared to be at the inflection point on the learning curve. At best, this method may tend to underestimate slightly the role of learning.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 6

FIDUCIAL LIMITS OF LEARNING

Indi- viduals	Administrations — Average of 6 Tests							Limits
	II	III	IV	V	VI	VII	VIII	10%
A	519	541	559*	576	592	594	583	54
B	573	547	564	602**	613	602	613**	38
C	569	582	597	636*	621	625	611**	58
D	543	581	599	635*	662	673	732**	74
E	493	544*	526	626	593	569	617**	48
F	628	648	664	667*	684	685	673**	39
G	566	612	653*	666	677	651	684**	49
H	521	598*	618	613	640	725*	666	55
I	518	561	578*	580	591	598	583*	51
J	600	612	630	633*	631	629	655**	32
K	537	571	562	592*	599	626	616**	45
L	550	605*	589	605	620	630	652**	45
M	541	553	556	555**	554	563	569**	43
N	590	596	586	610**	627	623	632*	38
O	467	481	525*	532	564	549	650**	50
T = 12/15				T = 1/15				

Key. * = point at which difference from score on administration 2 of 6 becomes as great as fiducial limits.

** = no difference as great as fiducial limits from administration 2 of 6 through given point.

from function to function. (2) and (3) are, of course, sources of intra-individual variation. However, since it was not possible in the present instance to differentiate satisfactorily the various factors contributing to the estimate of error,¹⁵ it was thought most parsimonious to exclude them from consideration as separate entities assignable to either intra- or inter-individual sources.

The Results—It should be stated at the outset that these results will be based on only seven administrations of each of the various tests, the first administrations being discarded as unreliable. They will also include only six tests in the primary analysis, since the norms for the card sorting test were found to be unsatisfactory.

Since the present sample is necessarily small, it is interesting to note these evidences of its representativeness. First, the

¹⁵The number of cases would be rather too small to give the curve-fitting methods much significance

INTRA-INDIVIDUAL DIFFERENCES

average standard deviation of the experimental group was over 90 per cent as large as that of the unmatched and unselected norm group. Second, the mean scores of the experimental group for administration number two are practically identical with those of the norm group. Three, the sample was split to allow an estimate of its internal consistency. Table 7 shows the result. This sort of consistency is surely one evidence of representativeness. These facts, combined, suggest the adequacy of the sample.

TABLE 7

CONSISTENCY OF SAMPLE

Factor	G_1	G_2	
I.D.	69%	75%	Analysis I
T.D.	76	77	Analysis II
R.V.	7	7	
Err.	17	16	

(Average proportions of total variation from the analyses of both series)
 G_1 and G_2 = respective halves of sample.

A fundamental assumption in the application of the analysis of variance is that experimental error is distributed with uniform though unknown variance about a mean of zero. Accordingly, Nayer's test¹⁰ for homogeneity of variance was applied to these data to check this hypothesis. In all cases, the value of L failed to reach even the 5 per cent level, which means that the variances within groups are the same. Statistically, this result justifies the application of the proposed method to these data.

Before turning to the results proper, it should be noted that they have a dual methodological aspect. Actually two separate problems exist; one is a problem of determining significance, and the second a problem of estimating magnitudes. The second problem ceases to exist if the first is not satisfied. In accordance with this fact, the discussion of significance will precede that of estimation in what follows.

¹⁰P. N. Nayer, "An Investigation into the Application of Neyman and Pearson's L Test, with Tables of Percentage Limits," *Statistical Research Memoirs*, I (1936), 38-56.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 8

INDIVIDUAL DIFFERENCES—SUMMARY

Factor	Degrees of Freedom	Sum of Squares	Mean Square	"F"	P	%
T.V.	618	5,373,669.90				
I.D.	78	3,713,633.72	47,610.69	23.43	< .01	72
R.V.	72	709,019.91	9,847.50			
Err.	468	951,016.27	2,032.09			

TABLE 9

TRAIT DIFFERENCES—SUMMARY

Factor	Degrees of Freedom	Sum of Squares	Mean Square	"F"	P	%
T.V.	615	7,148,910.67				
T.D.	75	5,488,873.51	73,581.62	33.68	< .01	77
R.V.	90	682,345.47	7,581.62	3.49	< .01	7
Err.	450	977,691.67	2,172.65			16

Tables 8 and 9 show the summary statistics for the two major series of analyses. These statistics were derived by adding the sums of squares and degrees of freedom from the separate analyses of the type illustrated in Tables 1 and 2. It should be observed that each of the factors is significant above the 1 per cent level, and that the per cent column contains the best available estimate of the respective magnitudes. These per cents are derived from the mean squares for the respective factors, minus the error mean square, and divided by a correction for the number of scores contributing to the means involved. Irwin¹⁷ has, in general, described the method.

A question may, however, be raised as to the validity of this procedure of adding sums of squares and degrees of freedom from the separate analyses of each series on the assumption that all the deviations are, in effect, grouped about a common grand mean. Although the tests were equated

¹⁷O. J. Irwin, "Mathematical Theorems Involved in the Analyses of Variance," *Journal of Royal Statistical Society*, XCIV (1931), 284-300 (especially pp 293-296)

with this criticism in mind, the question may best be answered by demonstrating that the same result is obtained if a method demanding no such assumption is employed.

First, it should be pointed out that the individual differences and trait differences factors are highly significant in each of the separate analyses in which they occur. Their total would, therefore, of necessity be highly significant. The repetitive variations factor, however, is *not* invariably significant in these separate analyses (cf. Table 2). Its total has, nevertheless, been shown by the method of adding sums of squares and degrees of freedom to be significant above the 1 per cent level. It is, then, this specific result which requires verification. Fisher¹⁸ has described a method appropriate for pooling the information from mutually exclusive though similar experiments. His technique makes it possible to sum the independent probabilities which arise from independent experiments by utilizing the fact that the log of the probability to the base "e" is equal to minus one-half Chi-squared. Two degrees of freedom are allowed for each independent comparison or probability value; these degrees of freedom and the log values are additive. The total may be tested for significance directly by entering the Chi-squared tables with the appropriate number of degrees of freedom. The hypothesis in the present instance is that the repetitive variations and error factors are from the same population. The obtained value of Chi-squared is highly significant and refutes this hypothesis as shown in Table 10. The table also shows that the results obtained by this method and by the method of totaling sums of squares and degrees of freedom are comparable, since by either procedure the obtained probability value exceeds the 1 per cent level by approximately the same amount.

Two problems, then, remain. The first concerns the determination of the significance of the difference between the mean magnitudes of individual and trait differences. The second relates to the relative importance of learning in the

¹⁸R. A. Fisher, *Statistical Methods for Research Workers* (London, Oliver and Boyd, 1936, sixth edition), 104-106.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 10

COMPARISON OF METHODS

R.V. = F = 3.49*	
Err.	ratio = 2.5
1% — F = 1.42	
ΣP ² s — X ² = 130.85	methods give equivalent results
1% — X ² = 50.89	ratio = 2.6

*Values taken from R.V./Err. in Table 9.

repetitive variations factor. The former problem was handled in the following manner. The per cent of the total variation contributed by individual differences was determined for each test (cf. Table 1). The per cent of the total variation contributed by trait differences was determined for each individual (cf. Table 2). These two series of per cents were then tested for the significance of the difference between their means. Since per cents are distributed as a binomial or Poisson distribution, it was necessary to transform the original measures before applying the test of significance.¹⁰ Fisher and Yates²⁰ have constructed a table of the inverse sine function which transforms proportions or per cents to angular degrees and normalizes their distribution. Fisher's "t" test was applied to these transformed values, and the obtained value of "t" was found to be insignificant at even the 50 per cent level. It was, thus, assumed that individual differences and trait differences were of comparable magnitude. This view that the tests are as discrete as the individuals—the specificity view of motor skills—is in substantial accord with the published conclusions of such investigators as Perrin, Muscio, Seashore,

¹⁰Clark and Leonard have contributed an excellent discussion of this point Cf. A. Clark and W. H. Leonard, "The Analysis of Variance with Special Reference to Data Expressed as Percentages," *Journal of American Society of Agronomy*, XXXI (1939), 55-66.

²⁰R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research* (London: Oliver and Boyd, 1938), p. 90.

Griffitts, and Buxton and Humphreys.²¹ It likewise agrees with the conception of motor abilities propounded by the authors of the "Minnesota Mechanical Ability Tests"²² These last make reference to the specificity view as "the theory of unique traits"

With respect to the latter problem, then, Table 9 shows that the repetitive variations factor accounts for approximately 7 per cent of the total variation in the series of analyses relative to trait differences. Table 5 shows that the initial (2-4) magnitude of the repetitive variations factor is approximately 10 times its final (6-8) magnitude. This suggests, as an estimate, that learning is at least 10 times as important a source of variation as are "random" fluctuations in the individual's efficiency from day to day in all functions.

Finally, it may be observed that if individual differences and trait differences are of comparable magnitude, and if the repetitive variations factor is of significant magnitude, then by definition intra-individual differences are greater than inter-individual differences. This fact was affirmed by determining the mean per cent contribution of individual differences to the total variation in the analyses of series one, and of trait differences plus repetitive variations to the total variation in the analyses of series two. The two series of per cents were appropriately transformed via the inverse sine function and the "t" test was applied to determine the significance of the difference between their means. The obtained value, confirming the hypothesis, was significant above the 1 per cent level.

²¹F. A. C. Pellin, "An Experimental Study of Motor Ability," *Journal of Experimental Psychology*, IV (1921), 24-56.

B. Muscio, "Motor Capacity with Special Reference to Vocational Guidance," *British Journal of Psychology*, XIII (1922), 152-184.

R. H. Seashore, "Individual Differences in Motor Skills," *Journal of General Psychology*, III (1930), 38-66.

C. H. Griffitts, "A Study of Some Motor Ability Tests," *Journal of Applied Psychology*, XV (1931), 109-125.

C. Buxton and L. G. Humphreys, "The Effect of Practice Upon Intercorrelations of Motor Skills," *Science*, LXXXI (1935), 441-442.

²²D. G. Pateison, R. M. Elliott, et al. *Minnesota Mechanical Ability Tests* (Minneapolis: University of Minnesota Press, 1930), p. 586.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

The Conclusions—In this population, and with respect to these functions, the following are the conclusions of the present investigation

- (1) Intra-individual differences were greater than inter-individual differences.
- (2) Individual differences and trait differences were of comparable magnitude.
- (3) Repetitive variations were of approximately one-seventh to one-eighth the magnitude of individual or trait differences.
- (4) Learning accounted for at least 90 per cent of the variation assigned to the repetitive variations factor.

NEW TESTS*

Test for Machinists and Machine Operators, by Joseph Tiffin, H. F. Owen, C. C. Stevason, H. G. McComb, and C. D. Hume. 1942. An achievement test of technical knowledge for machine shop operations. For 12th grade through adult level. Time, approximately 50 minutes. Machine or self-scored. Price, 18c per copy; specimen set 25c. Published by Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois.

The Purdue Pegboard, developed by the Purdue Research Foundation. 1942. A test of manual dexterity and facility for small assembly work. For high school through adult level. Time, two to four minutes. Price, \$9.75. Distributed by Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois.

Industrial Training Classification Test, Forms A and B, by Charles Lawshe and A. C. Moutoux. 1942. Discriminates between individuals likely to profit from industrial training programs and those likely to fail. For 12th grade through adult level. Time, 35 minutes. Price, 6c per copy; specimen set 15c. Published by Science Research Associates, 1700 Prairie Avenue, Chicago, Illinois.

Turse-Durost Shorthand Achievement Test, Form A, by Paul L. Turse and Walter N. Durost. 1942. Areas sampled are shorthand principles, shorthand penmanship or outline proportions, punctuation, paragraphing, sentence structure, and spelling. For first and second year shorthand students. Time, approximately 50 minutes. Price, \$1.10 per package of 25 tests; specimen set 15c. Published by the World Book Company, Yonkers-on-Hudson, New York.

The Behavior Cards, by Ralph M. Stogdill. 1941. Designed for use as individual test-interview with delinquent boys and girls. For ages 9 to 18. Time, 15 to 30 minutes. Price, \$2.50 per complete set, including specially constructed box, 150 cards, 25 record sheets, and manual of directions. Distributed by the Psychological Corporation, 522 Fifth Avenue, New York City.

*Prepared by Jane Gilbert.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Byrd Health Attitude Scale, by Oliver E. Byrd. 1941. Designed to measure health attitudes of the group or individual. For 11th grade level through college sophomore level. Time, approximately 30 minutes. Price, \$1.75 per package of 25 tests. Published by Stanford University Press, Stanford University, California.

Test on the Effects of War, by Lee J. Cronbach. 1942. A survey instrument, designed to study morale or confidence of high school youth. For high school students. Time, approximately 25 minutes. Price, 1c per re-usable test; 5c per answer sheet. Published by the State College of Washington, Pullman, Washington.

Traxler Silent Reading Test, Form 4, by Arthur E. Traxler. Revised, 1942. Includes rate of reading, story comprehension, word meaning, and power of comprehension. For grades 7 to 10. Time, approximately 50 minutes. Price, 7c per copy; specimen set 30c. Published by the Public School Publishing Company, Bloomington, Illinois.

MEASUREMENT ABSTRACTS*

Ackerman, Dorothy S. "The Critical Evaluation of the Viennese Tests as Applied to 200 New York Infants Six to Twelve Months Old." *Child Development*, XIII (1942), 41-53.

Bühler's Viennese Tests for the measurement of development of infants were given to 200 infants for the purpose of evaluating and validating them for use with American children. Representative groups of subjects were used. The procedure followed was that standardized for the tests. The average developmental quotient score was 106.67 as compared with a score of 100 obtained by Bühler for Viennese children. Split-test reliability coefficients ranging from .92 to .98, for the different age groups, were obtained. Suggestions for revising some of the items are made, but, on the whole, the test is considered to be a valuable, practical instrument for estimating the development of infants. *L. Bouthilet.*

Berger, A. "Test Construction and I.Q. Constancy." *Journal of Exceptional Children*, VIII (1942), 109-111.

Although much attention has been given to the effect of such factors as changes of environment, schooling, and glandular therapy upon I. Q. constancy, little emphasis has been placed upon the defects in the tests themselves as a source of inconstancy of the I. Q. This paper discusses some of the causes of I. Q. fluctuation. Among those listed are the fact that the I. Q. varies according to the particular test used to measure it, that the same test given at different age levels may involve the use of entirely different types of items, and that the variability of the groups upon which the tests were standardized may have been different. *L. Bouthilet.*

Berger, Arthur and Speevack, Morris. "An Analysis of the Range of Testing and Scattering Among Retarded Children on Form M of the Revised Stanford-Binet Scale" *Journal of Educational Psychology*, XXXIII (1942), 72-75.

The authors have found that a large percentage of retarded pupils increase their scores on the average 3.14 months of mental age when the tests are extended. The rhyme, digits forward and reversed, word naming, sentence memory (year XI), response to picture (Messenger Boy), and problems of fact are among the items most frequently passed beyond the first zero point. Frequent passing of certain items after a

*Edited by Forrest A. Kingsbury.

year's level of complete failures indicates the possibility of inadequate scaling for these items. It is suggested that the test should be extended at least to the point where two levels of failures have been reached, if it is to be an adequate measure in the clinical examination of retarded children. *Louise T. Grossnickle.*

Burt, C. *The Factors of the Mind: An Introduction to Factor Analysis in Psychology.* London, Univ. London Press. 1940 pp. xiv+509.

The Factors of the Mind reviews the field of factor analysis, particularly the English versions.

The logical methods rather than the results of factor analysis are discussed in the first section. The primary object of factorial methods is neither interpretation, which was Spearman's original concern, nor statistical prediction, which was Thomson's original concern. The object is description. "Mathematically, a factor is simply an average . . . of certain measurements empirically obtained. Logically, it is simply a principle of classification—a principle by which both tests (or traits) and the persons tested may be classified."

The section following describes the similarities among the various types of factor techniques. The last section is "an actual application of . . . the problem of temperamental types." The inverted factor technique is given its most complete review in this section. An appendix contains working methods and tables for computers. *Helen Wolfe.*

Canady, H. G., Buxton, C. and Gilliland, A. R. "A Scale for the Measurement of the Social Environment of Negro Youth." *Journal of Negro Education*, XI (1942), 4-13.

Seventeen environmental factors (social contacts, cultural, educational, home, etc.) considered by competent judges important for the mental development of Negro youth of high-school age are incorporated into an hour's interview. The subject's response on each factor is rated by the interviewer from 1 to 5 with the aid of a scoring key, yielding a total possible score ranging from 17 to 85. The *Environmental Inventory* items are less of the socio-economic type than those in the Sims scale (with which it correlates $.73 \pm .04$); and it has some relation ($r = .32 \pm .06$) to intelligence, as compared with the Sims scale correlation with intelligence, which was found to be $.16 \pm .05$. *F. A. Kingsbury.*

Carter, H. D. "How Reliable are the Common Measures of Difficulty and Validity of Objective Test Items?" *Journal of Psychology*, XIII (1942), 31-39

MEASUREMENT ABSTRACTS

Subject-matter tests taken by 200 psychology students were analyzed to determine the relative reliability of various measures upon which item selection may be based. Results indicated that accurate measures of item difficulty may be obtained from a representative group of as few as 25 students. The common measure of the power of test items to discriminate between good and poor students yielded a reliability coefficient of .46. The author concludes that a test may be improved more easily by basing selection on a measure of difficulty than on a measure of discrimination power. *L. Birdsall.*

Crissy, William J. E. and Flanagan, John C. "A Plan for Using Punched Cards in Presenting Test Results in Profile Form" *Journal of Applied Psychology*, XXVI (1942), 94-105.

The importance of keeping test results in profile form is urged by psychologists, counselors, and personnel officers. The authors have developed a method for graphic presentation of test results of the National Teacher Examinations of the American Council on Education, including a maximum of fifteen scores for each examinee. Scores on the tests used are reported on a common scale on which a specific score indicates a comparable degree of excellence for any one of the various tests. They are reported quickly, inexpensively, and in convenient form for permanent filing. The procedures for punching and interpreting the profile card are given in the appendix. *K. S. Yum.*

Dearborn, Walter F. and Rothney, John W. M. *Predicting the Child's Development*. Cambridge, Sci-Art Publishers 1941. pp. 360.

This report is based on the Harvard Growth Study, an investigation of physical and mental growth. Numerous tests and statistical procedures have been applied in an effort to determine constancy or variability of growth in intelligence, educational achievement, body size, ossification, and other characteristics. *Jane Gilbert.*

Deemer, Walter L. "A Method of Estimating Accuracy of Test Scoring." *Psychometrika*, VII (1942), 65-73.

When errors of test scoring obey a Poisson frequency law (theoretical considerations suggest that they do), the method described may be used for finding the upper fiducial limits of scoring errors per paper. A criterion is suggested for establishing tolerance limits on scoring errors, and a method is given (1) for finding the probability of being wrong in the statement that the tolerance limit is being met for a given

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

size sample or (2) for finding the size of sample that will make this probability not greater than some fixed value (Courtesy *Psychometrika*.)

Dodd, Stuart Carter. *Dimensions of Society*. New York, Macmillan, 1942. pp. 944.

This book presents a mathematical approach to society and represents an attempt to systematize statistical forms and data relative to society. The theory upon which this study is based is as follows "Any quantitatively recorded societal situation (S) can be expressed as a combination of time (T), space (L), a human population (P), and indicators (I) of their characteristics . . . each type analyzable into a specified number of *indices* each operationally developed by its exponents and each subdivided into a specified number of class intervals and further subdivided into a specified number of cases." While the field of sociology is emphasized in this presentation, the methodology should be applicable to all quantifiable data in each of the social sciences. *Jane Gilbert*.

Ezekiel, Mordecai. *Methods of Correlation Analysis*. New York, John Wiley and Sons. 1941. pp. 531.

Although this book treats statistical procedures largely from the economic point of view, it should be of general interest to measurement workers. It does not cover the entire field of statistics; rather, it deals with the types of relationships between variables. The author has attempted to bring up to date the interpretation of standard errors and to point out the application of the logical limitations to graphic curve flexibility. New and speedier methods of calculation and methods of estimating reliability of individual estimates are also presented. *Jane Gilbert*.

Feigelson, George A. "Item Selection by the Constant Process." *Psychometrika*, VII (1942), 19-29.

This paper relates the constant process used in psychophysics to the problem of item selection. Each test item may be described in terms of a limen, which is an index of the point at which an item discriminates, and the standard deviation of the limen, which is an index of the "goodness" of discrimination. The method developed may be related not only to the description of items but also to the description of persons. Thus a person's ability may be described in terms of a limen and its standard deviation (Courtesy *Psychometrika*.)

MEASUREMENT ABSTRACTS

Gillette, Annette L. "Relative Difficulty of Tests Within Each Year Level of Revised Stanford-Binet, Form L, Years Six Through Twelve." *Journal of Psychology*, XII (1941), 125-138.

"The data (from 506 cases) clearly indicate that within year levels there are variations in the difficulty of tests as measured by the percentage passing . . . The tables indicate the differences in difficulty of tests within levels and the reliability of these differences." Each of the 42 tests is named and numbered and placed in order of per cent passing of the total group. The tables will be of great value to the clinician. *Helen M. Wolfe.*

Greene, Harry A., Jorgenson, Albert N., and Gerberich, J. Raymond. *Measurement and Evaluation in the Elementary School*. New York, Longmans, Green, and Company. 1942. pp 639.

This book has been designed as a handbook of measurement for elementary school teachers and students of elementary education. Particular attention has been given to the problems involved in the construction, use, improvement, and interpretation of teacher-made examinations and tests. Important changes and trends in curriculum organization, instructional techniques, and in measurement and evaluation techniques have been incorporated in this edition, which is a revision of an earlier text.

The authors discuss types of educational and mental tests, the criteria of a good examination, construction and use of various types of standardized tests, the nature and use of intelligence and personality tests, measurement and remediation in specific academic areas, and finally, the use of test results for guidance purposes. While the book is directed at the elementary level, it should also be of general interest to measurement workers and teachers at all levels, particularly with reference to the discussions on test construction and standardization. *Jane Gilbert.*

Grossnickle, Louise T. "The Scaling of Test Scores by the Method of Paired Comparisons." *Psychometrika*, VII (1942), 43-64

The purpose of this study is to investigate, by the method of paired comparisons, a possible scaling of individuals who have made certain test scores, such that the additive property will be satisfied and such that a stability in scaling will be maintained—in other words, a scaling such that the scaled score of an individual will remain relatively the same regardless of the grouping of individuals in which he may be placed. The results show that it is possible to utilize psychophysical methods in psychological and educational test situations. Among the major findings are that Case V of the Law of Comparative Judgment is appli-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

cable to the data in this problem, the method of dividing the intermediate category equally between the greater and the less was the best of three possible methods, internal consistency was satisfied, and, finally, when a new test of stability was applied, it was found that the distances between the hypothetical individuals remain the same. (Courtesy *Psychometrika*.)

Guest, L. P. "Last vs. Usual Purchase Questions." *Journal of Applied Psychology*, XXVI (1942), 180-86.

The use of the questionnaire in market research has led to increased interest in the problem of the form of the question to enable the respondent to answer as easily and correctly as possible. The problem is to determine the difference between the two forms, "last purchase" and "usual purchase," in the questionnaire. The writer had two groups consisting of 438 college students, representing "last purchase" and "usual purchase" groups. Each student was asked to answer a questionnaire of 24 questions. The results show that the two questions give comparable answers for the most part when the results are treated for groups rather than individuals. In measuring the bland preferences, trends could be established equally well by either one of the two forms. K. S. Yum.

Guilford, J. P. "A Simple Scoring Weight for Test Items and Its Reliability." *Psychometrika*, VI (1941), 367-74.

It is pointed out that the scoring weights for test items should be approximations to regression-equation weights. For this reason any estimate of reliability of the weight should not be permitted to influence the size of the weight but should be used in determining the limit of acceptability of an item. A simple approximation weight is recommended for general use, and an *abac* is provided for the estimation of it when the correlation between item and criterion is the phi coefficient. A formula for the standard error of this weight is derived and tables of significant and very significant weights are presented in terms of deviation from the median weight. (Courtesy *Psychometrika*.)

Helson, Harry. "Multiple-Variable Analysis of Factors Affecting Lightness and Saturation." *American Journal of Psychology*, LV (1942), 46-57.

Factors affecting judgments of lightness (brightness) and saturation were evaluated through the use of analysis of variance. Judgments were made on an eleven-point scale running from zero to 10 for each attribute.

MEASUREMENT ABSTRACTS

All computations are shown and explained in detail. Judgments of saturation were significantly affected by background (white, gray, or black), intensity of illumination, hue, and the interaction of hue and background. Background was most important. Judgments of lightness were significantly affected by background, intensity of illumination, and hue. Illumination was most important. *Helen M. Wolfe.*

Holliday, Frank. "A Survey of an Investigation into the Selection of Apprentices for the Engineering Industry." *Occupational Psychology*, XVI (1942), 1-19.

The use of a battery of intelligence and aptitude tests improved the selection of English trade and engineer apprentices. Improvement was shown by a decreasing number of failures on national examinations, by foremen's satisfaction with the greater aptitude of their new apprentices, and by studies of the correlations between test scores and later success. Intelligence scores correlated with later success in mathematics, and aptitude scores with success in drawing. High intelligence scores alone were insufficient in predicting either the good trade or the good engineer apprentices. *Helen M. Wolfe.*

Holzinger, Karl J. and Harman, Harry H. *Factor Analysis*. Chicago, Univ. Chicago Press. 1941. pp. 417.

This book has been written to present the various approaches to the problem of factor analysis. The analytic and geometric bases for factor analysis are discussed as well as the theoretical development of various types of solution. Numerous practical illustrations are cited together with complete calculations. *Jane Gilbert.*

Jurgensen, Clifford E. "A Two-Dimensional Rating Scale." *American Journal of Psychology*, LV (1942), 255-60.

A two-dimensional rating scale developed for use in a boys' camp consists of ten traits or questions, each of which forms a scale representing five types of behavior. The first and fifth are apparently two opposite terms descriptive of the same trait; the middle or third is normal or average; and the second and fourth supposedly fall between the average and the extremes. The second dimension indicates the frequency of each type of behavior in terms of seven different degrees, such as constantly, almost always, usually, frequently, sometimes, hardly ever, and never. Administration of the scale and the scoring system are described. *K. S. Yum.*

Katz, Daniel. "Psychological Tasks in the Measurement of Public Opinion." *Journal of Consulting Psychology*, VI (1942), 59-65.

Polling opinion is useful as background for any successful campaign for influencing people. In fact, it is basic to the democratic process. In addition to such practical utility, it is important in the development of the science of social psychology. Many of the significant problems of social psychology which are difficult to handle in the laboratory can be profitably approached through the field study which ascertains attitudes and opinions. The author reviews the existing organizations and agencies as well as their type of work, and describes fundamental training and equipment for this field of public service. *Louise T Gross-nickle.*

Kent, G. H. "Emergency Battery of One-Minute Tests." *Journal of Psychology*, XIII (1942), 141-157.

A battery of brief tests is given, suitable for use as a preliminary measure in psychiatric examinations or under conditions in which presentation of longer, more formal tests is not feasible. Five oral tests and seven written tests are described. Some of the tests have not been standardized, and others are revised forms of previously published tests. The value of the tests for use in the military situation is emphasized. *L. Bouthilet.*

Link, Henry C., and Freiburg, A. D. "The Problem of Validity vs. Reliability in Public Opinion Polls." *Public Opinion Quarterly*, VI (1942), 87-98.

In spite of the fact that public opinion polls have attained the status of a scientific instrument and issues of national and international importance are being considered with reference to poll results, their use and interpretation are subject to error. In order to make possible an evaluation of reliability, a statement of the size and distribution of the sample of population interviewed should accompany each poll. High reliability, however, does not insure validity. One important check on the dangerous tendency to accept poll results uncritically has been their validation by periodic elections returns. Validations by comparison with specific purchasing behavior is also feasible. Questions on public attitudes and action should be framed in specific and behavioral terms rather than in general, stereotyped language. A discussion of various other practical techniques of validation is given, with the conclusion that the basic criterion of validity is behavior. *L. Bouthilet.*

Marble, Samuel D. "A Performance Basis for Employee Evaluation." *Personnel*, XVIII (1942), 217-226

MEASUREMENT ABSTRACTS

Better efficiency ratings can be secured from rating scales when their items deal with actual behavior on the job rather than with personality traits. After the descriptions of the job items are secured, they are evaluated. The relative importance of each behavior item to the job in question can be obtained by the psycho-physical method of equal-appearing intervals. Only items on which there is agreement among the judges are included in the final scale. Such a scale encourages the supervising officer to distinguish between the descriptive and evaluative function, and makes his task more palatable. *Helen M. Wolfe.*

Marshall, M. V. "A Study of the Stanford Scientific Aptitude Test." *Occupations*, XX (1942), 433-434

The test was administered to 47 students at the end of their sophomore year or the beginning of the junior year. Scores were then correlated, by the product-moment method, with the average science grade in the freshman and sophomore years, average science grade in the junior and senior years, and the average chemistry, physics, and biology grades for all four years, respectively. Twenty-five students took the test twice, once at the end of the sophomore year and again during the senior year. The results show that the test possesses high reliability but rather low validity. The author feels, therefore, that its practical utility with college students for the purpose of vocational guidance is open to question. *K. S. Yum.*

McNemar, Quinn. "On the Number of Factors." *Psychometrika*, VII (1942), 9-18.

A proposed criterion for the number of factors is developed on the basis of the similarity between a factorial residual and the partial correlation coefficient, something is known concerning the sampling error of the latter. Instead of computing the residuals as partials, a formula is presented for adjusting the standard deviation of the distribution of residuals so as to approximate the S.D. of the residuals as partial correlations. The criterion requires that factors be extracted until the adjusted S. D. reaches or falls below $1/\sqrt{N}$. When tried out on six samples drawn from six universes of known factorial description, the criterion indicated the correct number of factors each time. The requisites of situations adequate for such empirical checks are discussed (Courtesy *Psychometrika*.)

McQuitty, Louis L. "Conditions Affecting the Validity of Personality Inventories I, II; III." *Journal of Social Psychology*, XV (1942), 32-52.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

These three articles deal with conditions affecting the validity of personality inventories. The method of the study is to compare certain conditions affecting personality inventories with analogous conditions for intelligence tests as to nature of test items, content, directions to subjects, and the scoring of item responses; as to techniques of test construction, interrelations of item scores or answers, the selection and elimination of items, and scaling and scoring; finally, as to the nature of individual differences as influenced by both hereditary and environmental conditions. The author suggests possible ways of increasing the validity of the personality inventories. K. S. Yum.

Owens, W. A., Jr. "A New Technic in Studying the Effects of Practice upon Individual Differences." *Journal of Experimental Psychology*, XXX (1942), 180-183

R. A. Fisher's analysis of variance is suggested as a technique for obtaining an estimate of the effects of practice upon individual differences. The technique was applied to a study of motor skill tests, with individual differences and test administrations used as criteria of classification. The results show a tendency for individual differences to increase slightly, but statistically insignificantly, with practice. This tendency to remain constant suggests the importance of the initial selection program in industry. George W. Boguslavsky.

Reid, Scerley. "Respondents and Non-respondents to Mail Questionnaires." *Educational Research Bulletin*, XXI (1942), 87-96.

The accuracy of mail questionnaire results is difficult to estimate because of partial responses. In a study of the use of radios in Ohio schools an analysis was made of the difference in replies between first respondents, those responding to a follow-up letter, and a sample of the remaining group who responded only after intensive persuasion. Statistically significant differences were found between the groups, demonstrating that if the replies of the first group, or even the first and second groups together, had been used, erroneous and inaccurate conclusions would have ensued. Implications of the study for other investigations of the same type are that follow-up methods are necessary, that a representative sample of the non-respondents may be used to indicate the trend of their answers, and that in cases in which a follow-up questionnaire cannot be employed, the possibility of error must be recognized. L. Bouthilet.

Rodeheaver, Newton and Grim, Paul R. "Tests in Civics and Citizenship, Part II." *Social Education*, VI (1942), 222-224.

MEASUREMENT ABSTRACTS

This is the second installment of a bibliography of tests of various aspects of knowledge and attitude in the field of government. General headings include tests on the Declaration of Independence, the United States Constitution, community affairs, current affairs, and attitudes and beliefs. The objectives of the test, school grades for which it is suited, and a critical comment accompany each title. *L. Bouthulet.*

Slater, P. "Notes on Testing Groups of Young Children." *Occupational Psychology*, XVI (1942), 31-38

The basic principle of securing rapport and constancy of testing conditions among different groups of subjects, especially among young children of different ages, is a very important one. The author is particularly concerned with some of the conditions for the administration of the N.I.I.P. Group Test 70 on groups of children who are 11, but not yet 12 years old, and who are 13, but not yet 14 years old, respectively. In administering the test, the psychologist should consider the particular age group he is testing to secure the psychological condition of clear understanding and to meet the types of difficulty that are likely to arise. *Louise Grossnickle.*

Swineford, Frances. "Some Comparisons of the Multiple-Factor and the Bi-Factor Methods of Analysis." *Psychometrika*, VI (1941), 375-82.

Bi-factor and multiple-factor analyses of the same data are compared in two respects. First, two criteria are suggested for determining when the factorization is adequate. This problem being more acute for the centroid method than for the bi-factor method, the latter is used primarily for comparison only. It is shown also that the omission from the simple structure of entries smaller than .10 yields a pattern which is a poorer fit to the original correlations than is the bi-factor pattern. Second, the second-order general factor obtained from the intercorrelations of the primaries is found to be highly correlated with the general factor of the bi-factor pattern (Courtesy *Psychometrika*.)

Toops, Herbert A. "Code Numbers as a Means of Scoring Group-Administered Performance Test Products." *Journal of Applied Psychology*, XXVI (1942), 136-50.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Among the chief obstacles to the establishment of an adequate performance test program for guidance are the time and skill involved in the cost of scoring, and the delay between subsequent tests administered using the same equipment. In view of the fact that all mechanical performance test products have as a common feature space arrangements of movable sub-parts of a whole, and consequently exist in only a limited number of ways or patterns of correct and partially correct products, the author suggests the employment of "Addends" as a means of quick and certain identification of such performances. He shows in detail how to apply this addend principle by illustrations of fish-pole assembly and a bolt-and-washer assembly. *K. S. Yum.*

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Volume II

OCTOBER, 1942

Number 4

TWO ANNOUNCEMENTS	331
APTITUDE TESTS FOR ARMY WEATHER OBSERVER STUDENTS	335
<i>Earle Cleveland, Richard W. Faubion, and Thomas W. Harrell</i>	
THE OPTIMUM USE OF TEST DATA	339
<i>Maurice Lorr and Ralph K. Meister</i>	
A TECHNIQUE FOR TESTING UNDERSTANDING OF THE VISUAL ARTS	349
<i>Melvin W. Barnes</i>	
SOME OF THE LESS MEASURABLE OUTCOMES OF EDUCATION	353
<i>Edwin J. Brown</i>	
THE AIMS, OBJECTIVES, AND OUTCOMES OF THE OHIO TESTING PROGRAM	361
<i>Ray G. Wood</i>	
EDUCATIONAL REQUIREMENTS AND OCCUPATIONAL LEVELS	371
<i>Richard D. Allen and Lester F. Krone</i>	
THE PREDICTION OF SUCCESS OF STUDENT ASSISTANTS IN COLLEGE LIBRARY WORK	379
<i>Grace M. Oberheim</i>	
THE ADMINISTRATION OF GROUP TESTS	387
<i>Ernest M. Ligon</i>	
THE PURPOSE, ORIGIN, PLAN OF PROCEDURE, AND VALUES OF THE NATION-WIDE EVERY PUPIL SCHOLARSHIP TESTS.	401
<i>H. E. Schrammel</i>	
A TEST FOR SELECTING AND TRAINING INDUSTRIAL TYPISTS	409
<i>Clifford E. Jurgensen</i>	
MEASUREMENT ABSTRACTS	427
INDEX FOR VOLUME II	iii

Copyright, 1942, by
SCIENCE RESEARCH ASSOCIATES

STATEMENT OF THE OWNERSHIP, MANAGEMENT, CIRCULATION, ETC., REQUIRED BY THE
ACTS OF CONGRESS OF AUGUST 24, 1912, AND MARCH 3, 1933
OF EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT,
Published Quarterly at Chicago, Ill., for October 1, 1942

State of Illinois } ss
County of Cook }

Before me, a Notary Public, in and for the State and County aforesaid, personally appeared Walter A. Symons, who, having been duly sworn according to law, deposes and says that he is the Business Manager of the Educational and Psychological Measurement and that the following is to the best of his knowledge and belief, a true statement of the ownership, management (and if a daily paper, the circulation), etc., of the aforesaid publication for the date shown in the above caption, required by the Act of August 21, 1912, as amended by the Act of March 3, 1933, embodied in section 537, Postal Laws and Regulations, printed on the reverse of this form, to-wit:

1. That the names and addresses of the publisher, editor, managing editor, and business managers are Publisher, Science Research Associates, 1700 Prairie Avenue, Chicago; Editor, G. Frederic Kuder, 1700 Prairie Avenue, Chicago; Managing Editor, John H. Yale, 1700 Prairie Avenue, Chicago, Business Manager, Walter A. Symons, 1700 Prairie Avenue, Chicago.

2. That the owner is (If owned by a corporation its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding one per cent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a firm, company, or other unincorporated concern, its name and address, as well as those of each individual member, must be given.) Ralph A. Bard, 208 S. LaSalle St., Chicago, Ill.; Charles S. Boyd, Appleton United Paper Co. Appleton, Wis.; R. W. Glasner, 6409 W. 65th St., Chicago, Ill.; Alfred B. Hamill, 208 S. LaSalle St., Chicago, Ill.; Robert C. McNamara, 623 S. Wabash Ave. Chicago, Ill.; John T. Shaw, 135 S. LaSalle St., Chicago, Ill.; Lyle M. Spencer, 1700 Prairie Ave., Chicago, Ill.; Mrs. Dorothy Bard, c/o Roy E. Bard 134 S. LaSalle St., Chicago, Ill.; Roy E. Bard, 131 S. LaSalle St., Chicago, Ill.; George A. Bard II, c/o Ralph A. Bard, 208 S. LaSalle St., Chicago, Ill.; Miss Janet Bard, c/o Ralph A. Bard 208 S. LaSalle St., Chicago, Ill.; Robert K. Burns, 1700 Prairie Ave., Chicago, Ill.; Miss Grace M. Wagner, c/o Richard Wegner, 135 S. LaSalle St., Chicago, Ill.; W. C. Winkel, c/o Modins Mfg. Co., Starline, Wis.

3. That the known bondholders, mortgagees, and other security holders owning or holding 1 per cent or more of total amount of bonds, mortgages, or their securities are (If there are none, so state) None.

4. That the two paragraphs next above, giving the names of the owners, stockholders, and security holders, if any, contain not only the list of stockholders and security holders as they appear upon the books of the company but also, in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation, the name of the person or corporation for whom such trustee is acting is given; also that the said two paragraphs contain statements embracing affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees, hold stock and securities in a capacity other than that of a bona fide owner, and this affiant has no reason to believe that any other person, association, or corporation has any interest direct or indirect in the said stock, bonds, or other securities than as so stated by him.

5. That the average number of copies of each issue of this publication sold or distributed, through the mails or otherwise, to paid subscribers during the twelve months preceding the date shown above is (Not a daily publication.) (This information is required from daily publications only.)

WALTER A. SYMONS, Business Manager

Sworn to and subscribed before me this 21st day of October, 1942

DOROTHY MOEHLE, Notary Public

(SEAL)

(My commission expires June 12, 1946)

TWO ANNOUNCEMENTS

Although *Educational and Psychological Measurement* is several months short of celebrating its second anniversary, the growth it has made in its brief life is notable. Now it is possible to announce that another step forward has been taken. With this issue, *Educational and Psychological Measurement* for the first time goes to the members of the American College Personnel Association as their official journal. That this arrangement will result in a strengthening and broadening of the journal goes without saying.

Educational and Psychological Measurement will continue to serve the whole field of measurement as applied in education, industry, and government, and the pages of the journal will continue to be open to contributions from the entire field. In the past a number of outstanding articles have been contributed by members of the American College Personnel Association, although there was no tangible relation between the Association and the journal. It is a source of satisfaction that beginning with the January, 1943, issue the Association will be represented regularly by contributions from its membership in accordance with the new arrangement. A section on news of the Association will also appear in future issues.

An announcement to the members of the American College Personnel Association from its president follows.

G. FREDERIC KUDER, Editor

*To the Members of the American College Personnel
Association:*

It is with a great deal of assurance regarding the future of

our Association that I announce that an almost unanimous ballot approving the Executive Council's recommendation regarding our affiliations with this magazine has been received. The tremendous pressures that have been building up on our many members with regard to war work have caused your Executive Council to spend considerable time thinking of ways of fortifying our Association during the war period to the end that our corporate existence would continue. With the unusually fine co-operation of the editorial board of this magazine and of Science Research Associates we can be assured of more frequent contact with each other.

As president of your Association I want publicly to acknowledge the splendid work done by our Secretary, Dr. Feder, who conceived and initiated the plan for affiliation with this magazine as our official publication. Now Dr. Feder can go to war with the satisfaction of a job well done.

Unfortunately, the ballot voting was not completed in time for materials to be prepared for this issue. In future issues, however, articles and news notes will be presented. Grace E. Manson, Director of Personnel Research, Northwestern University, Evanston, Illinois, has agreed to act as editor for the A. C. P. A. section of the journal. All of us owe Dr. Manson a debt of gratitude for undertaking this service to our organization. I hope that each of you will feel responsible for making suggestions to this editor with regard to desirable materials to be included. News notes should also be sent to her. Please feel that your Executive Council wishes to do everything possible to further the work of our individual members and the continued strength of our Association.

Your voting was almost unanimous in favor of a restricted meeting in St. Louis in February. It appears that other personnel associations will follow our lead. We hope that this

will mean that our program, though restricted, may be a significant one. The Proceedings in abbreviated form most likely will appear in this journal.

As each of you finds yourself applying your personnel skills to the war work, may I urge you to keep clearly in mind that our contributions to higher education are needed more than ever these days for two reasons. First, all that college life adds to the maturing of students is as necessary in wartime as in peacetime. Our contribution through counseling of college students is being more clearly seen as a significant part of higher education. Second, we need to strengthen our techniques and our Association in preparation for the after-war period when tremendously increased enrollments may present the colleges with rich opportunities for affecting in a greatly increased manner the welfare of war students.

Your Executive Council extends to each of you a most cordial congratulation on your significant contributions to the war and wishes each of you continued effectiveness.

Cordially yours,

E G. WILLIAMSON, *President*
American College Personnel Association

APTITUDE TESTS FOR ARMY WEATHER OBSERVER STUDENTS¹

EARLE CLEVELAND and CAPTAIN RICHARD W. FAUBION
Army Air Forces, Technical Training Command

and

THOMAS W. HARRELL
University of Illinois

TWO groups of weather observer students in the Army Air Forces Technical Schools have been studied to find those tests which would be most predictive of success in the course.

The first group of students, numbering 116, entered the course in August 1940; the second, numbering 73, entered in November 1940. These students, like others described in previous studies² of selection procedures in the Army Air Forces Technical Schools, were selected on the basis of their being high-school graduates, with a score on a revised form of Alpha equivalent to a percentile rank of 75 and with a minimally acceptable score on a shop mathematics test.

The criteria with which the prediction test scores were compared are two grades in the weather observer course: the first based on a meteorology examination given about three weeks after the beginning of the course and the second being the final average for the three-month course. The grade on the meteorology examination correlated .70 with the final course average. The weather observer course covered the following topics:

1. Wind-aloft charts
2. Atmospheric soundings

¹A paper read at the 1941 meeting of the Midwestern Psychological Association.

²W. Harrell and R. Faubion. "Selection Tests for Aviation Mechanics," *Journal of Consulting Psychology*, IV (1940), 104-105.

W. Harrell and R. Faubion. "Primary Mental Abilities and Aviation Maintenance Courses" *Educational and Psychological Measurement*, I (1941), 59-66

3. Upper air observations
4. Plotting map signals
5. Surface observations
6. Weather forms
7. Weather instruments

After the content of the course had been studied, a tentative series of tests was chosen. These tests were:

- (1) *Mental Alertness*. An adaptation of the Henmon-Nelson Test for high-school students in which some of the items have been changed.
- (2) *Scattered X's*. A measure of perceptual speed in which the problem is to cross out the x's placed at random on a page of pied type.
- (3) *Identical Numbers*. A measure of perceptual speed in which the problem is to select which numbers in a column are identical with the number at the top of the column.
- (4) *Algebra*. A standard test of algebra.
- (5) *Meteorological Achievement*. 50 true-false items based upon material used in the Weather Observer course (This test is not to be confused with the meteorology examination, which is one of the criteria.)
- (6) *Physics Achievement*. 144 true-false items based upon material used in the Weather Observer course.
- (7) *Surface Development*. Six problems, each with six parts, in which a picture and a diagram of a simple object are shown, the problem being to match corresponding parts of the picture and the diagram.
- (8) *Flags*. 48 items in which the problem consists of deciding whether pairs of pictures of flags represent the same or opposite faces of the flags.
- (9) *Mechanical Movements*. 22 problems based on pictures of various mechanical movements as, for instance, a question about the direction in which oil will be forced, based upon a picture of the gears of a rotary oil pump.
- (10) *Cubes*. 32 problems in each of which the task is to distinguish whether or not two drawings represent the same cube turned to different positions.

Tests 2, 3, 7, 8, 9, and 10 are taken from Dr. L. L. Thurstone's Primary Mental Abilities study and were used with his permission. Test 4 was also used with Dr. Thurstone's permission.

APTITUDE TESTS FOR ARMY WEATHER OBSERVER STUDENTS

For the students entering in August, the means and the standard deviations of each of the ten prediction tests and of the two criteria as well as the zero-order correlations between the predictor variables and the criteria are shown in Table 1.

TABLE 1

MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR TEST SCORES AND GRADES FOR 116 W. O. STUDENTS

Variables	Mean	Standard Deviation	Correlation with Meteorology Exam	Correlation with Final Course Average
Meteorology Examination	68.5	16.8	—	.70
Final Course Average	80.6	4.9	.70	—
Mental Alertness	70.0	10.8	.39	.39
Scattered X's	26.2	9.2	.11	.11
Identical Numbers	50.6	6.3	.28	.32
Algebra	3.1	2.3	.47	.41
Meteorological Achievement	10.8	7.3	.40	.40
Physics Achievement	33.9	24.0	.55	.45
Surface Development	21.4	6.8	.18	.11
Flags	24.5	10.7	.16	.10
Mechanical Movements	22.7	12.0	.41	.32
Cubes	15.8	6.8	.17	.17

It will be noted that, in spite of the previous selection of the subjects by means of a test of mental ability, the mental alertness test correlated .39 with each of the two criteria. The meteorological achievement test, devised by the Classification Division, A.A.F.T.T.C., to measure meteorological concepts, correlated .40 with the grade on the meteorology examination. The physics achievement test and the algebra test correlated .55 and .47, respectively, with the same criterion. Other correlations between test scores and the grade in meteorology were positive but not so high. The Surface Development Test, which has consistently correlated significantly with grades in the basic mechanical course at Air Corps Technical Schools, correlated positively but insignificantly with the Weather Observer course grades, which is not inconsistent with what one would expect.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

A multiple correlation between the grade on the meteorology examination and the best combination of the tests resulted in a correlation coefficient of .63. The tests included were mental alertness, meteorological achievement, physics achievement, and algebra. A combination of the mental alertness, the meteorological achievement, and the physics achievement tests yielded a multiple correlation coefficient of .62. The difference between the two coefficients was not enough to warrant the additional testing time necessary for the algebra test. The regression equation was

$$\bar{X}_0 = .29 X_1 + .38 X_2 + 30.34,$$

where \bar{X}_0 = the *most probable* meteorology examination grade,

X_1 = the meteorology aptitude test score (meteorological achievement plus physics achievement),

X_2 = the mental alertness test score.

This regression equation, based on the August class, was used to predict the results for the class entering in November. Of the 73 November students, 10 were eliminated before the completion of the course; of these 10, 8 fell below a critical level of 55, calculated from the regression equation. Of the 63 students who completed the course, nine passed who were predicted to fail.

Conclusions:

(1) Evidence is given from a cross-validation study that an examination made up of mental alertness, meteorology, and physics questions significantly improves the selection of weather observer students in the Army Air Forces Technical Schools.

(2) One of the tests, Surface Development, which has been shown to be predictive of basic airplane mechanics grades, does not correlate significantly with weather observer grades. This suggests the specificity of requirements for the various training courses within the Army Air Forces Technical Schools and indicates that a single selection procedure is inefficient.

THE OPTIMUM USE OF TEST DATA¹

MAURICE LORR

and

RALPH K. MEISTER

THE procedures conventionally adopted in administering and scoring age scales of the Binet type are often wasteful of time and test materials. For many practical situations, a more economical procedure is much in need. It is the purpose of this paper to describe a briefer method of administering and scoring age scales, to indicate the several advantages of the newer method, and to present comparisons of the results of this newer method applied to the Revised Stanford Binet with results from the conventional form of the scale and from the abbreviated scales.

The rationale for the method to be described is directly derived from the fundamental relationships between the field of mental test theory and the methods of psychophysics, in particular the constant method of psychophysics. Let us first briefly review the procedure in determining a sensory threshold such as the two-point tactual limen by the constant method. An appropriate range of stimuli that are judged neither "two" nor "one" 100 per cent of the time is selected. Each stimulus is then administered to the subject by means of the aesthesiometer a large number of times in a prearranged order. The subject judges the presence or absence of the desired experience, which is "two." The responses are then classified and the relative frequencies of judgments of "two" and "one" for each grade of the stimulus scale are determined. The limen, which

¹The authors wish to express their thanks to Dr. Martin L. Reymert, Director of the Mooseheart Laboratory for Child Research, for his generous permission to use data from its files.

may be computed by the constant process, represents a transition zone between stimuli too weak to arouse a response of "two" and stimuli strong enough to elicit a response of "two." Conventionally the stimulus value that elicits a response of "two" 50 per cent of the time is regarded as the stimulus limen.

Now let us consider in the mental age scale the groups of items, supposedly equal in difficulty, that are allocated to each year level as representing the typical performance of individuals of the corresponding chronological age. These items are such that the response is classed as either "correct" or "incorrect." Each item may be regarded as having a characteristic response-value that differentiates "incorrect" from "correct" responses, and thus each item requires for a "correct" response a given degree of ability as expressed in terms of a certain age group. This response-value corresponds to the stimulus value of psychophysical discrimination. Theoretically if an individual were presented with items ordered as to difficulty and if his responses were made without measuring error, correct responses would be made up to a certain point on the scale depending upon the individual's ability. "Incorrect" responses would be made to all items beyond this point, and the scale value of this point—which corresponds to the psychophysical limen—would represent a measure of the individual's ability or intelligence.

Actually of course, as with sensory thresholds, no such point exists. In actual practice, instead of this sharp theoretical division we obtain mixed successes and failures over a number of year levels. It has been pointed out (1) that such irregularity of performance or scatter is in part a consequence of the lack of perfect correlation between items resulting from a lack of homogeneity and from the presence of error. In the mental test situation, therefore, it is seen that the response process may be regarded as a composite consisting of a characteristic or "true" component of ability and an error component. When the ability component of the individual plus the chance error component of his response is greater than the

level of ability required to pass the item, he answers correctly; when this composite is less, he answers incorrectly. The discrepancy between the actual level of the response and the assumed true value, of course, constitutes the error.

It is evident from these considerations that the psychophysical method of constant stimuli for determining stimulus thresholds is applicable to such mental test data. Thus, when items are arranged in order of difficulty for standard age groups, and the response of any individual to any item can take only two values such as "correct" or "incorrect," the frequency distribution of responses as a function of item difficulty may be assumed to be the integral of the normal probability curve. The characteristic response-value or test limen of that individual (his mental age score equivalent) will be that difficulty value expressed in terms of age that yields "correct" responses fifty per cent of the time. The individual's variability or error will be the standard deviation of the probability function described (2).

Thus, by simply computing a test limen for an individual in terms of the age level at which he passes 50 per cent of the items, we have an alternative method of determining mental age. This procedure is much shorter than that required for the full scales and even shorter than that required for the abbreviated scales. The test limen or mental age is determined either by (a) the single age level at which the individual passes 50 per cent of the items, or by (b) simply interpolating for the 50 per cent point which falls between the age level at which he has passed more than 50 per cent and the next higher level where he has passed less. Linear interpolation is justifiable here since in the range concerned the curve of per cent passing is practically linear, all of the data will ordinarily be employed, and a measure of individual scatter is not desired. The limen or mental age score may be computed by linear interpolation by the following formula:

$$\text{M.A.} = a_m + \frac{(a_1 - a_m) (.50 - p_1)}{(p_m - p_1)},$$

where a_m -- the age at which more than 50 per cent of the items were passed.

a_l the age at which less than 50 per cent of the items were passed.

p_m -- the per cent of correct responses at a_m .

p_l the per cent of correct responses at a_l .

The procedure of the examination itself is as follows: Begin testing at the age level where the child is likely to pass half and fail half of the items. On the average this age level will be within six months of the child's chronological age. The examiner should, of course, also take into account the grade placement, general behavior, and any additional facts available concerning the child's ability. If the child responds correctly to 50 per cent of the items at the age level where testing is begun, his test limen or mental age score is exactly that year level, and the test is completed. Should the child respond correctly to *more* than 50 per cent of the items, the items at the next higher (older) level are administered. Testing is continued until 50 per cent of the items or *less* are passed, and in most cases only one additional level is required. In only a few cases is more than one additional level of testing required. This is what might be expected from an assumption of a normal distribution of intelligence in the general population.

Should the subject respond correctly to *less* than 50 per cent of the items at the level where testing is begun, the items at the next lower (younger) level are administered and the examination is continued until a point is reached where the subject passes either 50 per cent or *more* than 50 per cent of the items. It should be noted that one level determines the test limen or at most two. If the subject has been tested through three or four levels before a limen determination is possible, as may sometimes happen, the only data that are to be used in determining his mental age score are the level at which he passes 50 per cent of the items or the two adjacent levels at which he passes more and less than 50 per cent of the items, respectively. The rest of the test data is ignored.

THE OPTIMUM USE OF TEST DATA

As Terman has stated, mental ages beyond fifteen are artificial and are to be regarded as simply numerical scores. It was decided by the present authors to express test limens beyond the age of fourteen in terms of these artificial mental ages instead of chronological age. The individual's test limen is, therefore, simply the mental age level at which 50 per cent of the items are passed. The problem arises as to what mental ages should be assigned to the Average Adult and Superior Adult I, II, and III levels. At lower age levels, such as fourteen, a child passing half of the items at that level is credited with a mental age of fourteen by the liminal method. Similarly, individuals passing half of the test items at the upper levels should be assigned the following mental age scores:

Average Adult	15 years, 4 months
Superior Adult I	17 years, 4 months
Superior Adult II	19 years, 10 months
Superior Adult III	22 years, 10 months

MENTAL AGE INTERPOLATION TABLES GIVING THE NUMBER OF MONTHS CREDIT TO BE ADDED TO THE LOWER AGE LEVEL TO DETERMINE A GIVEN MENTAL AGE FOR A CHILD

1.

		For half-year levels from II thru V		
		0	1	2
Items passed at lower age level	4	2	2	3
	5	2	3	4
	6	3	4	5

2.

		For levels from V thru XIV		
		0	1	2
Items passed at lower age level	4	3	4	6
	5	5	6	8
	6	6	7	9

3.

		Items passed at Average Adult level			
		0	1	2	3
Items passed at age XIV	4	4	5	6	9
	5	6	8	9	12
	6	8	9	11	13

4.

		Items passed at Superior Adult I		
		0	1	2
Items passed at Aver. Adult	5	5	6	10
	6	8	10	14
	7	10	13	17
	8	12	14	18

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

5.

		Items passed at Superior Adult II		
		0	1	2
Items passed at S. A. I	4	8	10	15
	5	12	15	20
	6	15	18	23

6.

		Items passed at Superior Adult III		
		0	1	2
Items passed at S. A. II	4	9	12	18
	5	12	18	24
	6	18	22	27

Tables have been prepared to facilitate the process of interpolating for the test limen when it falls between the lower (younger) age level at which the child has passed more than half of the items, and the next higher (older) age level where he has passed less. The number of items passed at the lower age level is given at the left and the number at the higher age level is given at the top of each table. The body of each table gives the number of months of credit (to the nearest whole number) to be added to the age corresponding to the lower age level. For instance, if a child passes five items at age seven and two items at age eight, we enter Table 2 at the left and the top, to find that the second row and the third column intersect at the value 8. Thus to the lower age level of seven is added eight months to yield a mental age of seven years and eight months. Table 1 is to be used for determining limens that fall within the age range, II through V, where each half-year is regarded as a separate age level. A table for interpolating between IV-6 and V is unnecessary since its values are the same as those in Table 1. Table 2 is to be used for finding limens within the age range, V through XIV. In each instance the number of items passed below the limen, i.e., at the lower age level, is found at the left of the table. Tables 3, 4, 5, and 6 are self-explanatory. There will, of course, be a few instances in which the liminal method and the process of interpolation are impossible, as for example, when four items are passed at Superior Adult III level.

In order to compare the liminally determined mental ages with those conventionally computed on the full and abbreviated

THE OPTIMUM USE OF TEST DATA

scale, one hundred Revised Stanford Binet test folders were chosen at random from the younger age group in the current files of the *Mooseheart Laboratory for Child Research*. Tests chosen were restricted to those of children from approximately seven to eleven in order to avoid limen determinations at the adult levels where the changed scoring rationale would obscure the basic comparison desired. Successes and failures at each age level in the administration of the full scale were recorded on cards together with the C.A., M.A., and I.Q. for that test. Then, by a consideration of only those items which are part of the abbreviated scales, each test was rescored and a M.A. and an I.Q. calculated for an *assumed* abbreviated scale administration of the same test. Then these same tests were rescored a second time and assigned a mental age score determined by the test limen as described above, and a corresponding I.Q.

Assuming the full scale as the standard (and disregarding the form, L or M), the abbreviated scale M.A. scores and the test limen mental age scores were each correlated with the M.A.'s of the full scale. The correlation coefficients were .98 for the abbreviated scales and .91 for the test limen method. Then the I.Q.'s corresponding to these ages were correlated with the I.Q.'s on the full scale. The coefficient of correlation between I.Q.'s on the full and abbreviated scales was .97 and that between I.Q.'s on the full and the test limen scales was .83. The correspondence of scores may be further judged from the fact that the mean absolute discrepancy between the ratings on the long form and on the liminal form was a little less than 7 points. In 75% of the cases the discrepancy was less than 10 points; in 91% less than 15 points; and in 97% less than twenty.

The savings in time of testing may be appreciated from the fact that for these one hundred cases the average number of levels administered in the full scale was 6.6 while the average for the test limen determination was only 2.2 or roughly a third. Even when the abbreviated scales are used and only four out of six tests at each level are given, the

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

number of tests is cut only one third. Thus, the average number of tests administered in the abbreviated scales for this sample is still twice that required for the test limen determination.

In order to determine how this cutting of the length of the test affected its reliability, out of the one hundred tests originally selected 31 pairs were chosen which represented successive administrations to the same individual child. In this way data for reliability calculations upon a test-retest basis were secured. It should be noted, however, that between test and retest there was an interval of elapsed time of about one year and that these reliability coefficients might be expected to be lower than those of the usual test and retest immediately following because of the changes in the individual and the conditions of testing and even the change in the content of the test itself since items of the same degree of difficulty were not administered the second time. The reliability of the I.Q. score for the long form of the test for the 31 cases was .86, for the abbreviated scale .75, and for the limen form .61. This drop in reliability with the use of fewer and fewer items for the determination of the score is roughly in accordance with the expectations derived from the Spearman-Brown prophecy formula.

The equivalence of the mental age scores secured by the two methods of scoring may also be judged upon the basis of the statistics presented below.

Form	I. Q. Mean	I. Q. Standard Deviation	M. A. Mean	M. A. Standard Deviation
Full	108.1	16.1	113.53	22.5
Abbreviated	106.4	15.7	114.92	22.6
Liminal	108.9	14.0	114.37	19.6

The I.Q. and the M.A. means are practically the same for the three sets of scores. The variability of the I.Q.'s and the M.A.'s naturally decreases as we pass from the full form to

abbreviated and liminal forms because of the reduction in the length of the test.

In the evaluation, on the basis of these data, of the limen method of scoring as opposed to the simple addition of number right, it should be pointed out that this estimate of its effectiveness is specific to the Revised Stanford Binet and is affected by the extent to which the Revised Stanford Binet satisfies or does not satisfy the conditions for a true difficulty scale. Though the test limen method of scoring assumes a series of items ordered in difficulty, this condition is only approximately met in the Revised Stanford Binet in the sense that though items at age ten are invariably more difficult than items at age five, yet between adjacent age levels there are inversions, as a number of empirical studies have shown. Such a condition might have been expected from the limitations imposed upon the placement of items. In general they had to be placed according to difficulty. In addition, the demands of variety, interest, etc., had to be satisfied for each age level. It is quite probable that if an age scale homogeneous as to content and rigidly ordered as to difficulty were obtainable, the liminal method would provide the most reliable measure of an individual's performance on a limited number of items. This would be the case since the individual is scored upon the basis of his performance on items that are of 50 per cent difficulty for him.

It will be a matter of concern to some examiners that there is no spread of performance to be analyzed. Or perhaps they will feel the need of some measure of individual variability. However, as it has been pointed out (1), the practice of inspectional analysis of individual successes and failures to secure a crude estimate of the individual's "primary" abilities is at best questionable. Such scattering of passes and failures is based for the most part on factors inherent in the test, in test construction, and in systematic errors. Furthermore, measures of individual variability on the Revised Stanford Binet possess no unique significance for the individual (3).

The abbreviated method of constant stimuli thus enables an examiner to secure mental age scores reasonably equivalent to those obtained by the conventional method of scoring of the full scale, in half the time ordinarily required. Of course, when administration time is ample, the full Stanford Binet should be given. However, occasions arise in the clinic and in the field when time is at a premium. At such times use of the shorter method enables the examiner to administer a test that would not otherwise be possible at all.

REFERENCES

1. Lorr, Maurice and Meister, Ralph K. "The Concept of Scatter in the Light of Mental Test Theory," *Educational and Psychological Measurement*, 1 (1941), 303-310.
2. Mosier, Charles L. "Psychophysics and Mental Test Theory: Fundamental Postulates and Elementary Theorems," *Psychological Review*, XLVII (1940), 355-366.
3. McNemar, Quin. *The Revision of the Stanford Binet Scale* (With an introductory chapter by L. M. Terman). New York: Houghton Mifflin, 1942, 185.

A TECHNIQUE FOR TESTING UNDERSTANDING OF THE VISUAL ARTS

MELVIN W. BARNES

University of Illinois

IN 1941-1942, the University of Illinois offered a survey course in Literature and Fine Arts. This course, open to sophomores, is one of seven full-year courses which now comprise the lower-college program of the General Division of the College of Liberal Arts and Sciences. One unit of the course in Literature and Fine Arts is devoted to the study of painting. In an attempt to appraise achievement in this phase of the year's work a testing technique was evolved which it is the purpose of this note to describe.

In the conduct of the course the students were brought into contact with a wide variety of paintings by means of lantern slides. These works of art were studied in terms of a four-fold scheme of analysis: color, composition, expression, and function. In some instances, a work was studied first in black and white for the purpose of emphasizing composition before it was considered in color. On a number of occasions two or more representations of a single theme or incident by different artists were studied comparatively. Since the basic aim of the course was to cultivate understanding and thereby—it was hoped—appreciation, little time was spent on artists, history, or techniques of painting. In addition to the reproductions used in the classroom, materials owned by the Department of Art and others in the University museums were made available to the students. The course was rounded out with a day at the Chicago Art Institute.

When the problem of testing achievement in this course arose, the following procedure was devised. So far as the writer knows, the device is unique.

The technique employed two projecting lanterns to throw simultaneously two paintings in color on adjacent screens placed in the front of the classroom. By this method colored reproductions of a size approximately three feet by five were placed side by side in a way which permitted every member of the class to see them clearly. The paintings had not been seen by the class before the time of the test. The test, which was mimeographed, was based upon points of similarity and contrast between the paintings thus reproduced. Before the showing of the paintings each student was given a copy of the test and time was allowed for reading the directions. By the adjustment of Venetian blinds the room was darkened enough to permit good vision of the projected pictures, while enough light was admitted for reading and writing. Of those taking the test the only writing required was the indication of responses by a letter written in a blank space.

The pair of paintings was selected chiefly for their numerous points of contrast. One painting was a Rubens, the other a modern work by William Gropper. The test items were organized in accord with the scheme of analysis and synthesis which had been followed in class. The first set of items dealt with color, the second set with composition, and so on. A variety of the conventional multiple-choice items was used. The following is an example:

In Painting A (the Rubens was designated A) more use is made of $\left\{ \begin{array}{l} \text{a. linear contrasts} \\ \text{b. straight lines} \\ \text{c. sharp angles} \\ \text{d. rhythmic curves} \end{array} \right\}$ than in Painting B.

The test is in process of analysis, the results of which will be used in revising and lengthening it.

This technique obviously does not require any particular type of test item but is adaptable to all of the conventionally used forms involving comparison and contrast. Reproductions of works of sculpture and architecture could, of course, be utilized as well as those of painting. Since this method affords

TECHNIQUE FOR TESTING UNDERSTANDING OF VISUAL ARTS

a means of providing colored reproductions which are precisely relevant to the aims, content, and method of a course, it appears to have possibilities for classroom use which the standard tests on the market do not possess. This type of test, moreover, is decidedly inexpensive, whereas the cost of the better standard tests in art places serious restrictions upon their use.

SOME OF THE LESS MEASURABLE OUTCOMES OF EDUCATION*

EDWIN J. BROWN
Kansas State Teachers College

I NEED not say that I appear before this group with considerable apprehension and not a little hesitation. Frankly, it is not the group which is causing my trepidation, but my *subject*.

When one speaks of *outcomes* in education he is talking about the very essence of it all. He is talking about our end-product, the thing for which we throw in the current, gear up the machinery, put in the man-hours, spend the money; the thing which we get after we've done the work. Outcomes in education are for us what the finished interceptor, the eight-gun pursuit plane, the one-ton bomb, the 155-millimeter field piece, the thoroughly trained airman is to our war program. Thus my hesitation in discussing *outcomes* in education at all, even those which we are agreed are more or less measurable. To discuss the *more measurable outcomes* before this group would take some boldness and one should attempt it with much hesitation, but to discuss the *less measurable outcomes* is about twice as risky. My only comfort is that no one knows much more about it than does his neighbor. And I am not supposed to tell how to measure them.

First of all, what are some of the outcomes (may I assume there are such) which we want to get—outcomes which are difficult of measurement?

May I say that after spending some fifteen years rather directly in the field of measurement, I am not nearly so certain

*Paper read at the meeting of the National Association of Teachers of Educational Measurement, San Francisco, February 24, 1942.

of the efficiency of the work as I once was. Heresy? Right—but you can throw me out of the organization later. In general, I've about come to the conclusion (there are exceptions, of course) that the ease and accuracy with which any educational outcome is measured is in direct proportion to its unimportance. That is, the easy items to measure accurately are the ones which make the least difference whether they are measured or not. I agree that there are notable exceptions to my generalization. In general, though, you agree with me, don't you, that the more important things of life tend to be beneath the surface, too deep to be picked up readily on the hooks of a question, and that measurement is usually involved in questioning, either direct or indirect.

Does this mean we should not try to measure these things? I'd say, "Certainly not." Let's work on the thing rather than say that it is one of the unmeasurables and that we can't do anything about it.

First of all, I'd like to start with the thought that the more difficult of measurement outcomes fall into general classes. (There may be three, six, or nine.) Let us call two of these difficult-to-measure groups, for want of a better term at this time, *Outcomes in Attitudes* and *Outcomes in Appreciations*. I can measure fairly accurately some outcomes in arithmetic skills, in spelling accuracy, in verb usage, but I seem to have much difficulty in measuring the same youngsters in their *attitudes* toward arithmetic, toward spelling, toward grammar. I find myself relying on clues which are not too clearly defined in my own mind, when I try to measure their attitudes. Can these clues, then, be developed, expanded? Are the big things in life, after all, caught rather than taught? I sometimes get confused, and when so, am inclined to say yes. If we go into attitudes we can break them down into any number of divisions. There are attitudes toward school, toward home, toward boys, toward girls, toward law and order, and so on, exhaustively. However, these can be grouped into two big categories from which a further breakdown might

come I refer to attitudes and traits which are primarily concerned with *personal growth* and attitudes and traits primarily concerned with our relationship with *others*. Each general item is susceptible to a further breakdown, of course. This I shall suggest later.

Isn't one of the wrong assumptions we make when we speak of the *more unmeasurable* outcomes of education that we are inclined to fail to do what we always do in working with the more measurable items, viz., break them down into some of their component parts? We take arithmetic computation and break it down into the four fundamental operations of adding, subtracting, multiplying, and dividing. Then we take addition and break it down even further, trying to find not only the weakness in addition but the cause of that weakness. Might we not, if we tried seriously, break down attitudes concerned primarily with *our own personal growth*, into smaller units, breaking these in turn into still smaller ones until we might secure parts small enough to be measured? Of course, we would not be sure that an old axiom would not be ruined and we'd find after we did our measuring that the whole is not equal to the sum of the parts.

Suppose we take the topic of attitudes in *personal development*. What are some of the things which might be considered from the viewpoint of a high-school boy or girl? Of course no one can name all of the desirable attitudes which are worth considering, but let's begin:

1. An attitude of *open-mindedness*. We might interpret this as the Evaluative Criteria for the Cooperative Study for Secondary School Standards does. A willingness to revise opinions and conclusions in the light of new evidence.
2. An attitude of *critical-mindedness*. Disposition to seek causes or explanation, to weigh evidence with care, and to withhold judgments until sufficient evidence is in.
3. An attitude of *concentration*. Ability to give attention through a considerable period of time in spite of difficulties or distractions.

4. An attitude of *industriousness*. Disposition to use time and ability effectively and constructively.
5. An attitude of *responsibility*. Willingness to acknowledge responsibility for one's acts and obligations.
6. An attitude of *self-reliance*. Willingness to make decisions and carry out plans oneself instead of depending on others or the school.
7. An attitude toward *self-control*. Ability to avoid display of temper or other uncontrolled emotion.
8. An attitude of *creativity*. Desire to do or say things in a new or better way.
9. An attitude of *enthusiasm*. Readiness to enjoy life and participate in its wholesome activities.

These we will all grant have something to do with *personal development*.

There can be little doubt but that some of these items are more susceptible to objective measurement than are others. Again, there is little doubt but that each of these is more susceptible to measurement than is the general outcome used for illustrative purposes from which they came, the outcome of *personal development*.

My suggestion is now that each of the items be considered in turn for a further breakdown. This, of course, would entail the development of a valid definition, which probably would go back to the common consent, massed judgments technique.

In the field of *social relationships* we have another of the more difficult to measure outcomes of education. No one, of course, argues that the outcome is insignificant because it is hard to measure, or that it is found completely embodied in other outcomes which are easier to measure. Shall we then pass it up entirely? Let's see what we might do to it, again falling back on the *Cooperative Study* material for suggestions. We suggested that open-mindedness, critical-mindedness, concentration, industriousness, responsibility, self-reliance, self-control, creativity, and enthusiasm are desirable, but difficult to measure, outcomes of *personal development*. Now what are

some desirable outcomes, broken down just once, of *social relationships*? Suppose we say:

1. *Social-mindedness*. Willingness to subordinate personal advantage to the common good.
2. *Co-operation*. Desire to work agreeably with others.
3. *Tolerance*. Good will toward groups or individuals of different race, customs, or opinions.
4. *Courtesy*. Consideration of others.
5. *Generosity*. Willingness to share opportunities or privileges.
6. *Honesty*. Integrity in handling money, straightforwardness, sincerity in personal relationships.
7. *Dependability*. The extent to which one fulfills promises, discharges obligations, finishes tasks.
8. *Loyalty*. Devotion to interests of friends, school, home, country.
9. *Fair play*. Unwillingness to take advantage of others or another.

Our difficulty, of course, is to get measurements of attitudes toward, not just of information about. One can comparatively easily measure information about, but not the attitude toward.

One could go on and build these up. The point I would make is that while these outcomes are difficult of measurement, each breakdown tends to become more objective or perhaps better, *less subjective*. If in turn one were, for instance, to analyze honesty for a high-school pupil, it might be found that a fairly valid test could be set up. I'm inclined to guess that if the validity could be assured, reliability would follow fairly readily; that is, a test could be made which would agree with itself.

Appreciations is the generic name we give to another group of outcomes which are deliberately sought. We have, however, not gone so far as we might in developing measures, largely, I suppose again, because of the feeling of intangibility. Undoubtedly, these are even more difficult of measurement, as the emotional factor enters in. However, again, it is not out

of place to say that appreciation of *beauty in nature*, or better, *in art*, is not so difficult to measure as appreciation *in general*. Appreciation of *commendable conduct and qualities in others* might be measured—indifferently well perhaps—but still measured.

Appreciation of *home* and *family* would seem susceptible of some measurement, and so on for other items such as appreciation of good workmanship, appreciation of spiritual and religious values, appreciation of law and constituted authority, and others.

It seems to me that a prime reason for not doing more, at least in the attempt to measure certain educational outcomes, lies in our unwillingness to attempt a further breakdown which is always a step in measurement. To illustrate: Under outcomes of social relationships which are surely end-products of education, I listed among others, *courtesy* and *fair play*. Let's see what a further breakdown would do. We defined *courtesy* simply for the sake of mutual agreement as consideration for others. Let's start a further analysis, considering the subject from the viewpoint of a junior in high school, somewhat as follows: Do I always wait my turn; do I refrain from loud talking and laughing when it disturbs others? Do I refrain from interrupting others when they are talking? Do I offer to share what I have with others? And so on. Isn't it possible that this rather intangible outcome of education can be measured, and much more reliably than we are inclined to think?

Fair play, which is one step down from the general outcome, *social relationship*, might in like manner be susceptible to analysis. Fair play; unfair advantage; cheerful loser; modest winner; recognition, appreciation, and commendation of skill in others; consideration of the sensibilities of others; abiding by decisions without question either by word or act; loyalty to personal and team ideals; observation of training rules in athletics; winning without boasting, losing without whining; willingness to sacrifice for a group good, and so on. Each item in turn might be analyzed still further. We make

the mistake of trying to measure an important item like this with a short test. It requires a test as well developed as a Binet revision.

Perhaps, in conclusion, it should be said we must not be too much perturbed at first about the measurement of these outcomes but should turn our attention first to securing these outcomes with a greater degree of certainty. It would be desirable indeed to be able to measure results in a citizenship class—if we are sure that we are teaching citizenship. Patriotism is dear to every American's heart, but who knows *for dead sure* what it is—or how to teach it?

I would indeed be disconsolate did I not believe we are doing better work both in teaching these intangible outcomes as well as in measuring them than we think we are. The clues we get are probably fairly good. This is, of course, wishful thinking. Someone has said, " 'Tis better to travel hopefully than to arrive," and Browning puts it, "A man's reach must exceed his grasp or what's a heaven for?"

THE AIMS, OBJECTIVES, AND OUTCOMES OF THE OHIO TESTING PROGRAM*

RAY G. WOOD

Ohio State Department of Education

TODAY the world is moving along at an increased tempo and at a higher degree of efficiency. Education, too, must swing into step, and it is doing so. Education, it is agreed today, is for the whole of life; it is concerned with the development of the "*whole nature*" of every pupil," with the integration of his total personality for the good of himself and of society. There is disagreement, however, as to just how much time and attention the school should give to the development of the social, moral, and physical phases of the child, and how much time and attention to the development of his intellectual side.

Some progressive extremists would sacrifice the mental development of the student to the personalizing of his character, or make it a complementary factor only; traditional extremists would reverse the procedure. Neither policy is wholly applicable to our present educational situation or meets the needs of the typical American school.

What is needed now, more than ever, is that from our theorizing and experimentation some tangible and definitely constructive guiding principles, practicable in all our public schools, be evolved for the development of integrated social-intellectual personality. What these shall be is still a question, but I believe with many that we should train the mind to the maximum of possibility, taking into account the limitations of all methods, salvaging the best that is in them, and inventing new ones that are more generally successful.

*Paper read at the meeting of the National Association of Teachers of Educational Measurements, San Francisco, February 24, 1942.

In doing this we shall arrive at new methods, new objectives, new subjects, and new curricula. But to make possible this more comprehensive program, with its broad conception of the schools' activities, time (so essential to everything now) must be saved. Overlapping, useless, and minor material will have to be discarded to make way for the many newer and more important elements that are to be added; that is, a better, clearer understanding of the *tangibles* needed will leave more time for the needed intangibles to be developed.

And it is for this saving of time, it is for the determination of achievement in the needed tangibles, that a good testing program is requisite

We in the testing field in Ohio are justly proud that our work is helping to do this, that it is in line with the modern philosophy of education, in fact, that it is giving direction to it in a concrete and personal way. By providing the great majority of teachers in Ohio a scientific means of analyzing and evaluating their product, we are saving them time to achieve more, and more efficiently.

Our program, which has been carried on since 1929 as a division of the State Department of Education, is in its several phases unique to Ohio. Its chief objective is the motivation of scholarship—it stimulates the educational units to put forth more effort and seeks to increase the efficiency of that effort. There is no compulsion whatsoever about participation in our program, and there is no attempt at standardization. These are important features, we believe, contributing to its effectiveness and its popularity. Besides, it is distinctly a product of and for the Ohio schools, because the tests are built by Ohio teachers for Ohio children. They are designed not to determine *the success or failure of the individual but to help teachers to adjust* their teaching to the needs of their children and the children to adjust themselves to the work of the particular class or subject. Because of this, students in the state no longer approach tests with fear and trembling, in such a disturbed state emotionally and mentally that the results are impaired, but they anticipate the testing with a spirit of sports-

AIMS, OBJECTIVES, OUTCOMES OF THE OHIO TESTING PROGRAM

manship and with a realization that whatever the results may be, they will aid, not hinder, them

Since the beginning of our testing program, the democratic philosophy of education has been the guiding principle upon which the work of the Ohio Scholarship Tests division is based. The tests of the program are revised annually, and thus provision is made for meeting the changing curricula, textbooks, and methods of educational practice. Furthermore, there is no compulsion to use any of the tests administered. Schools, private as well as public, are free to use *any* phases of the program for such purposes as they wish.

Stated concisely, the objectives of the several phases of our program are:

1. To provide materials for the improvement of instruction.
2. To provide a continuous program of new and improved tests
3. To provide for the motivation of students toward greater accomplishments in their classroom activities.
4. To provide pertinent instructional research data.
5. To provide curriculum guides.

These objectives are achieved variously by the six distinct phases of our testing program, which are: *The Every Pupil Tests*, *the Eighth Year Test*, *the General Scholarship Test for High School Seniors*, *the District-State Scholarship Team Test*, *the Senior Survey Tests*, and *the Bulletins of Research and the Curriculum Guides*.

An idea of the popularity of the program may be gained from these figures: a total of 1,200,772 tests were administered in this program last year. In this number was represented every county in the state and over a thousand large and small city schools. Of this number, 41,269 were eighth-graders; 1,146,672 were grade-school and high-school pupils who took part in the Every Pupil Tests; 5,305 were high-school seniors in the upper third of their classes; and 7,526 were high-school students in grades 9 to 12, who were selected

to take part in the annual spring academic and commercial contests.

This popularity is due to the fact that the school men of the state look upon the testing program as one of the most vital and beneficial functions in the state educational set-up. They recognize the testing program as one of their own supervisory tools, as an active force growing out of and along with the actual conditions in their schools, not as a measuring stick imposed from without.

A brief discussion of the several phases of this program will give a clearer picture of the program.

The *Every Pupil Tests*, because they affect the greatest number of children, may be considered the most important phase of our program. There are two series of these tests. the *First Every Pupil Test*, administered in December for general diagnostic purposes; and the *Second Every Pupil Test*, administered in March for an achievement measurement and as a check upon the effectiveness of the remedial teaching which has been carried out on the basis of the results of the *First Every Pupil Tests*.

In these series are included tests for all the subjects that are most commonly taught in Grades 3 through 12. For example, we have tests in English for Grades 3 through 12; Reading for Grades 2 through 12; Mathematics for Grades 3 through 10; and in the other subjects such as Latin, French, Geography, Social Studies, Chemistry, Physics, General Science, Biology, Health Education, and Hygiene. New forms of each test are developed for each administration, and new tests are added from time to time, such as *Attitudes and Skills in the Use of References*, *Conservation*, and *Scientific Thinking*.

These tests are of the achievement type and are so constructed that they give good general diagnoses of both the individual and the class abilities and deficiencies in the preparation in the specific subject. For an analysis of the more puzzling and particular individual difficulties, the teacher must use specifically diagnostic and functional tests; but, for the

AIMS, OBJECTIVES, OUTCOMES OF THE OHIO TESTING PROGRAM

general group and the average pupil, these tests have proved very satisfactory in the thirteen years they have been administered.

As was mentioned before, these tests have their origin in the Ohio classrooms. They are constructed by Ohio teachers of recognized ability, working in committees or individually; they embody suggestions sent in by teachers and administrators throughout the state; and they are validated against research studies, committee reports, the most used textbooks, and the *Ohio Curriculum Guides*. In this way we are sure that the tests are really measuring *what is or should be taught*, and that they are of service to the teachers and students in emphasizing and vitalizing the important content.

Each subject-test takes one forty-five minute period and may be administered either in the individual classroom or as best suits the purposes of the school. All tests are scored by the classroom teacher. Forms are furnished with each order for recording for state use the general distribution of the scores of a class and for making item reports. These reports are, of course, kept in strictest confidence. As soon as they are received by the state office, state percentile and item norms are compiled from the data. These are then printed and mailed to the participating schools.

By analyzing and interpreting the results of the work of her own class in the light of these norms, the teacher is able to determine wherein there are deficiencies and can set about to determine the particular causes of the weaknesses. Similar analyses and interpretations may be made for the individual pupil—in fact in not a few instances the students themselves analyze and interpret their own results. Thus, comparatively early in the year both the teacher and the student have an estimate of the equipment of the student for the course and indications of probable areas of difficulty, and together they can set about to discover the causes of these difficulties. When the causes are determined, then a definite remedial program may be settled upon.

After the *Second Every Pupil Test* has been administered

and the analyses and interpretations have been made, a comparison with the results of the *First Every Pupil Test* reveals whether progress has been made and whether the remedial procedures have functioned effectively. Through these comparisons every pupil has a diagnosis of his achievement and has evidence of the phases of the subject he has mastered and the phases upon which he must still concentrate. This is the factor that is emphasized continually—the improvement of the original product. It has done much to lessen, if not to erase completely, the fears of teachers that these tests are a means of measuring teacher efficiency. Teachers have come to realize that high scores or low scores on the tests are not the important factors, that the really important factor is the improvement of learning as indicated by the raising of the scores of the individual pupils and of the class between the first and second series. They know that a low-ranking class that shows progress is a better evidence of good teaching than a high-ranking class that remains at the same level from test to test.

The *Eighth Year Test* and the *General Scholarship Test for High School Seniors* are two of the other important parts of the Ohio Scholarship Testing program. These tests are measures of the cumulative achievement of these respective groups and are administered in the spring of each year. The tests are designed to measure not only factual knowledge as such but also the ability to use this knowledge in functional situations, and to stimulate the desire for the acquisition of such knowledge and ability. The *Eighth Year Test* is a two-hour test in the four fields of English, mathematics, history, and science, and may be taken by any eighth-grader. The *Senior Test* consists of rather general tests in the following five areas—English, mathematics, science, social science, and reading and functional language, with 30 minutes allowed for each test. Because one of its chief purposes is the selection of outstanding students, only the upper third of the graduating seniors are admitted to this examination in Ohio. However, it is administered to *all* high-school graduates in the state of

New Mexico, where it is administered under the direction of the University of New Mexico at Albuquerque. Other colleges and universities outside Ohio likewise use this test in their freshman placement programs.

High general scholarship is evidenced not merely by an acquaintance with the basic principles in the general fields of learning but likewise by the ability to apply this knowledge to life situations; it calls for a broad as well as a thorough educational preparation. The objective in these two tests of our program is the stimulation of this high general scholarship, and it is evident that they are serving as real stimuli in this respect. Follow-up studies have proved that the specific incentives of these tests have resulted in an increased interest of students in their achievement, have encouraged many to broaden their educational preparation through more wide and general reading, and have led many high-school students to choose more widely from the courses offered in their program of studies. Hundreds of scholarships are awarded annually by many colleges and universities in and outside of Ohio to seniors ranking high in this test. Follow-up studies have shown that these students do considerably better than average work in the institutions at which they matriculate; and research has shown that these tests are predictive for the group as a whole of probable success in continued education.

The fourth part of our testing program—the *District-State Academic and Commercial Scholarship Tests* are administered in May of each year at the five state universities (in Ohio well located geographically for this purpose). The *District-State Academic Test* has become the "scholastic event" of the year in Ohio; it fosters interest in academic achievement in a manner akin to that in which athletic events stimulate athletic prowess. The students enter with a great deal of enthusiasm and interest, and they work hard in order to make the scholarship team and participate in this "academic field day." Last year 7,526 students in a total of 237 teams participated. Schools send teams of 32 students, or fewer, to

their closest university center to compete with other teams and students in that district for academic honors. Each team is limited to two entrants in each of sixteen subjects, which include English, mathematics, history, the sciences, Latin, and modern languages. The schools are classified according to enrollment, except that schools of a county system combine to send one team. Points are given for the first twenty places in each of the subjects, and the teams are ranked superior, excellent, or honorable mention, according to the number of points they accumulate. District awards are made to individuals and to teams according to the classifications of the schools. Then all papers are sent to our state office, where similar awards are computed for the all-state winners.

The purpose behind this part of the program is again a motivation of scholarship, especially by the granting of recognition to students of outstanding achievement. It should be noted that it is not only the thirty-two pupils on the team who receive this motivation, but also the hundreds of others who try to make the team. The truly professional and able administrator makes the most of this opportunity and encourages every student to strive to win a place on the scholarship team, by not announcing the appointees until just before the meet. We have athletic contests, music contests, and forensic contests; why shouldn't we have academic contests? Why shouldn't we popularize the "brains" as well as the "brawn" of our schools? Ohio schools have recognized the values of this academic competition, and school people would not be happy if it were to be discontinued.

The fifth phase of our program—*Senior Survey Tests* and remedial materials—are comprehensive diagnostic tests designed for locating deficiencies in the fundamentals in English usage, reading, and mathematics, and are administered the first week of either semester. Along with the tests are provided manuals and workbooks for the remedial work, which are so organized that the instruction may be carried on individually or in groups, with a minimum of direction on the part of the

teachers. This part of the program had its origin in the plea of college leaders who found many high-school graduates entering their institutions poorly equipped in these fundamentals so necessary to the carrying on of successful work. If high-school students who enter college are in need of remedial work in these areas, many who do not go on to college probably also need to have their ability in these fundamentals improved. Recognizing the urgency of this need, the Ohio State Department of Education is granting one-half unit of high-school credit to each senior who shows proficiency in overcoming the weaknesses clearly indicated by his results on Form A of these tests. Many colleges and universities throughout the nation have recognized the merit of these materials and are using them as a part of their freshman program of testing.

So much for the tests themselves and the valuable services they render to teacher and pupil alike, when properly administered, analyzed, and interpreted. Let me suggest briefly some of the concomitant values that are to be derived from such a co-operative testing program.

The first is the great number of research studies that the results of these tests give rise to and that are of particular value to the teachers of the system because the bases lie in the local situation. In Ohio many such studies have been completed, and teaching procedures have been influenced to the end that the indicated weaknesses are being remedied. Very complete studies have been made in English, mathematics, and the social sciences, and less comprehensive ones in other subjects. Two research reports, R₁ and R₁₀, have recently been issued on the *Every Pupil Tests*—these give superintendents and teachers techniques for determining growth learning curves on the basis of their own class scores, and of interpreting their significance to the individual pupil and teacher.

A second very important concomitant value is the aroused interest of teachers in the improvement of their classroom product and in an understanding of the scientific diagnosis and measurement of their teaching procedures. This has been

evidenced by the vitalizing of curricula materials and by the functionalizing of the learning process. Most of the tests for the entire program are built by individual classroom teachers whose work has been recognized as outstanding or by committees of teachers in the field of the subject. Teachers and administrators alike have expressed amazement at the broader understanding of their task and the other worth-while results that have come to them from their participation in these test-building projects. Not only have teachers taken an active part in the construction of the tests, but they have also been active in the writing of *Curriculum Guides*, which suggest materials to be taught and abilities to be developed in the various fields, recommend methods of procedure, and provide a working bibliography. These *Guides* are not attempts to dictate order in the presentation of material or methods of procedure nor are they attempts to supplant the local course of study; they are designed to assist teachers and local curriculum committees in the re-examination, re-evaluation, and re-formation of their curricula in the light of present trends in educational philosophy.

A third and most important concomitant value is the stimulation and motivation of the thousands of students who annually come in contact with this program. The *Every Pupil Tests* help them to help themselves; they have specific and objective evidence of their achievements and of their lack, of their abilities and of their deficiencies, and go about their remedial classwork with understanding and with determination to improve. The other tests create an active interest in not just good scholarship, but in excellence of achievement.

Time does not permit the listing of more of these values nor the further elaboration of those already mentioned—each in itself would furnish material for another paper. However, these suggestions and this brief survey of our Ohio Scholarship Testing Program have, I hope, presented the possibilities of a varied, comprehensive, and co-operative state testing program, and demonstrated that such a program is of real service to its participants—teachers and pupils alike.

EDUCATIONAL REQUIREMENTS AND OCCUPATIONAL LEVELS

RICHARD D. ALLEN and LESTER F. KRONE
Department of Public Schools, Providence, Rhode Island

IN every job description and worker description there appears the category "Educational requirements." These educational requirements for almost any kind of work are somewhat elastic. In good times when labor is scarce, standards always move downward; while in periods of unemployment, standards are automatically raised. This is true with the standards of college and professional schools as well as with the standards of apprenticeship and of employment in the less skilled classifications. In fact, most employers and personnel workers regard educational requirements as merely a convenient and economical screen with which to eliminate the less desirable applicants.

This situation is interesting in view of changes in educational practice during recent years. Formerly most school adjustments were made in terms of *grading*; that is, the slower pupils were kept back grade after grade until there was a high percentage of over-ageness for the grade throughout the school system, at least until the legal age of school-leaving began to operate. Under such conditions the "last grade attended" really had a definite meaning in terms of school achievement and educational qualifications. In recent years, however, there has been a strong tendency toward the practice of promoting most children from grade to grade largely on the basis of age and attendance. Under such conditions, differentiation of instruction has been accomplished by classification or grouping within the grade, or by group assignments within each class. Consequently the "*last grade attended*" by any child may not be a fair indication of his educational quali-

fications. In fact almost the only accurate method of appraisal of school achievement is by means of standardized tests, the results of which are relatively independent of the school system, the school, the curriculum, the teacher, and other factors such as the policies of pupil adjustment in the individual school or school system.

A study of achievement among pupils of any grade indicates a distribution of scores covering a range of from five to eight school grades or educational ages. Under these circumstances, a diploma or a grade of school leaving means little unless it is supplemented by such information as the teachers' marks in academic subjects, marks in special subjects, information regarding the curriculum, and the classification of the pupil, and especially, if possible, marks in standardized tests in the basic skills and core subjects. These matters are not generally understood by employers and personnel workers. Instead they usually condemn the school product in a rather general and wholesale fashion, assuming that schools still are like those with which they were once familiar. An illustration in point may be helpful at this stage:

In a large manufacturing plant the personnel manager, a very capable and wise man, called the school placement office for a bus boy to work in a lunch room, clearing away the dishes and carrying the trays. A strong body, a willing and pleasant disposition, and a reasonably agreeable personality were the chief requirements. There was little or no opportunity for advancement, and experience had shown that an intelligent boy would not remain long at such work. The counselor selected a big, strong, sixteen-year-old boy from a special class for backward children. He was nominally a seventh-grade pupil, but actually his achievements in arithmetic and reading were about on the third-grade level. However, he was the kind of boy that the teachers would always send on simple errands or use as a helper in routine tasks. He was cheerful, willing, clean, and from an average home background. The counselor explained all of this to the personnel

manager, who accepted the boy and reported after the first month that he was doing very well.

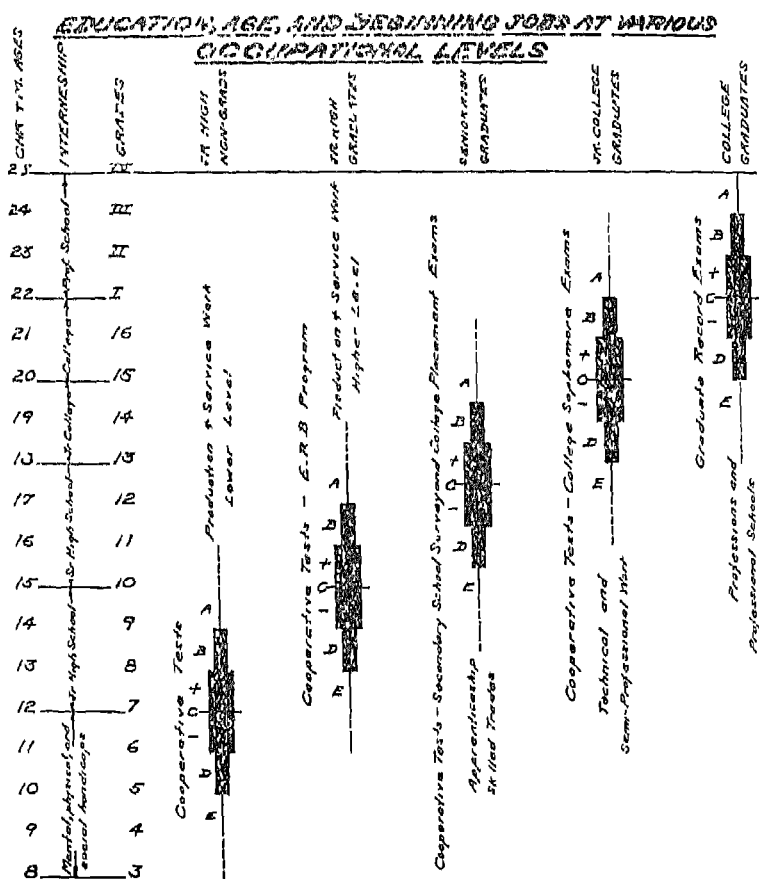
One day, however, the office boy in the main office left to enter the military service; a second boy was ill; and a third had been sent on a long errand. It was necessary to move a number of heavy articles from the office and the bus boy was requisitioned for the purpose. In the midst of his work the general manager called him and sent him on an errand. The directions would not have been difficult for *an average boy*, but they were extremely difficult for him. He had to ask directions and it became evident that he could not read the names of officials on the various doors and some of the other strange signs around the plant. When this was reported to the manager he roundly condemned his personnel man for hiring such a stupid boy, and the school placement office for recommending a boy who could not read, and the school system for graduating a pupil under such conditions. This man was evidently thinking of the schools of a generation ago, and of conditions in employment at a time when bright and capable high-school graduates were glad of an opportunity to work from the bottom up. I am sure that he did not mean to be unfair or unjust.

If personnel officers and counselors are to be realistic and truly helpful to young people, it is extremely important that they should neither over-estimate nor under-estimate their achievements and abilities. An accurate appraisal of educational status and achievement is absolutely necessary in order to determine the *readiness of any individual to enter a program of training at any occupational level*. While a rough approximation of status and achievement may be obtained from the school record, it is at present possible to bring any record up to date by means of a battery or inventory of achievement and aptitude tests. No single battery will adequately serve the entire range of individual differences to be found among adults. Instead at least five different batteries would seem necessary to accomplish the purpose. Even a few years ago it would have been difficult to have selected adequate

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

test batteries for this purpose, but at the present time there are at least five excellent batteries which are available and in common use, and there are at least five different groups or levels for which these tests may be appropriately used.

These levels in both educational and occupational opportunities are shown in the accompanying chart:



Interpretation: A person's present status may be estimated by his last grade or by his test record. In any occupational level, a person in the highest third (A or B) is preferred and can usually obtain work even in slack times. A person in the middle third can usually obtain work in normal times. Persons in the lowest third may meet the minimum requirements, but can improve their chances of getting and holding a job by entering at a lower level and progressing through in-service and supplementary training. This procedure solves most problems involving discrimination.

RICHARD D. ALLEN,
LESTER F. KRONE.

EDUCATIONAL REQUIREMENTS AND OCCUPATIONAL LEVELS

The lowest level shows a group of approximately two thousand pupils on a seventh-grade level in the Providence junior high schools. The figure shows the distribution of these pupils in a battery of achievement tests. Both the Metropolitan and the Co-operative tests have been used and have shown approximately the same range and distribution of scores. It is interesting to note that more than a thousand adults and young people over eighteen years of age, who applied for training as production workers in jewelry and novelty manufacture, had a median score in reading, general mathematics, and academic aptitude equal to the seventh grade in the junior high schools. For these jobs there were practically no requirements in regard to education. The work was repetitive and did not require even a mechanical aptitude or finger dexterity beyond the "D" level. However, all seven hundred graduates of the jewelry class were trained and placed after less than one month's training at a minimum wage of forty cents an hour.

The second figure is typical of the distribution of scores on the Co-operative tests at the end of the 9A or the beginning of the 10B grade. The median non-graduate school leaver in Providence has attained a tenth-grade status, has a C—I.Q., and ranks C in reading comprehension and general mathematics. It is interesting to note that in a group of more than two thousand men and women who were candidates for training in defense classes in Providence, the median grade of school leaving was the tenth, and the median scores in academic aptitude, reading comprehension, and general mathematics were approximately equal to the tenth-grade median. These jobs for the most part, in the opinion of employers, required ability to read directions and a reasonable mastery of fundamental skill in mathematics. These workers may be classified as production workers of the higher level.

The third figure indicates the results of the Co-operative tests on approximately two thousand high-school graduates in Providence, showing range and percentile distribution. From these pupils are selected those who are admitted to colleges,

technical schools, apprenticeship schools, and nurses' training schools, as well as many of the more desirable occupational opportunities which require high-school graduation.

The fourth figure indicates the range and distribution of scores on the college sophomore Co-operative tests as shown in the Co-operative testing program of both local colleges in this area and the nation-wide program. Persons at this grade level usually make their decisions for specialization in college or for entrance into higher skilled occupations on the technical and semi-professional level. For instance, such persons are preferred for officer training in the armed services

The fifth figure shows the range and distribution of scores on the college Graduate Record Examination and also among the candidates for teacher-training positions in the Providence schools on the National Teacher Examinations over a period of the past decade. At this level people are selected for entrance to the graduate schools and the professions

The arrangement of these five figures opposite a scale, showing chronological and mental ages at the left and school achievement in terms of grades at the right, is for the purpose of facilitating appraisal and comparison. The procedure is somewhat as follows: Indicate by a check mark in the first column the person's chronological age and by a cross his mental age. In the second column indicate by a check mark his last school grade and by a cross either his average scholarship rank as shown by his school record or the battery of achievement and aptitude tests at that level. This may be done by a horizontal comparison of values indicated in the appropriate figure. Then draw a horizontal line from the cross on the grade measure intersecting the figures representing occupational levels. Frequently this line will cross two or even three figures. If the line does not fall in the upper half of the figure it is possible, and even probable, that the worker will find it easier and more profitable to attempt to meet the requirements on the lower level in which his qualifications will give him a preferred status.

EDUCATIONAL REQUIREMENTS AND OCCUPATIONAL LEVELS

This chart is to be used only in determining the *most advantageous place of entrance into an occupational level. It helps to get a person on the pay roll on a realistic basis.* The second step in the process is to work out a program of supplementary and in-service training and experience which should help the individual to improve his educational and occupational status and thus to earn promotion, after which he may, if he so desires, transfer to a job in a higher occupational level.

The general use of objective methods in appraising the qualifications of applicants and the use of such devices as the present chart for purposes of comparing different levels as well as in determining the status of a person within any group, will be an important step in promoting equality of opportunity and preventing favoritism and group discriminations. For instance, a relief worker was assigned to a technical research project because he had attended a college for two years and expressed an interest in such work. He liked the prestige of his assignment and refused other kinds of employment despite the fact that no such jobs were available to one of his qualifications. Tests showed that his present educational skills were about those of the average eleventh-grade pupil and among the lowest 3 per cent of college sophomores. When shown his results on the chart by the counselor, he decided to enter a defense training class and now has a good job as a production worker. His social worker was also glad to have the support of objective data based upon the measurements of abilities of the worker rather than upon his observations or opinions.

A similar instance was that of a star athlete who graduated from a high school with the respect of both faculty and student body. He wanted to become an apprentice machinist and felt that he was rejected because he was a Negro. His school record showed C's and D's in most subjects. Moreover, these marks were obtained in non-college subjects and in the slow-learning class sections. His scores in aptitude tests and in the Cooperative tests in English, Social Studies, Mathematics, and Science all placed him in the lowest fifth of the

class. On the basis of such evidence he could be shown that many non-Negros would also be rejected even with much better qualifications. However, his qualifications would easily admit him to a defense training class and his excellent character and personality record, as well as his physical assets, would make him desirable as a production worker in the same plant where he had been rejected as an apprentice. Moreover his pay as a production worker would be much higher than as an apprentice. In addition, if he still wanted to become a machinist, he could enroll in free evening courses in mathematics, drawing, science, and machine shop practice, and when he had mastered the necessary fundamentals and skills, he could be examined and certified by the school authorities, by the state civil service, or by the state director of apprentice training, and on the basis of such evidence he should be able to obtain employment as a machinist. In recent years the placement office has had requests from a number of employers for at least a few Negros of superior qualifications to indicate that they have abandoned the practice of race discrimination. They were unwilling to employ them or to reject them *because* they were Negros, but would be glad to employ them if they could demonstrate that they were as well or better qualified than others who were being employed.

Accusations of prejudice and discrimination are the perpetual alibis of the unsuccessful candidate for any job. The only effective answer is the more general use of objective methods in determining the qualifications of candidates and more objective and accurate production records in determining promotions.

THE PREDICTION OF SUCCESS OF STUDENT ASSISTANTS IN COLLEGE LIBRARY WORK

GRACE M. OBERHEIM
Iowa State College Library

THERE are many problems which arise in connection with the selection and use of student assistants in college library work. The specific problem which led to this study was the difficulty of obtaining student assistants in the Loan Department of the Iowa State College Library who were capable of doing the required work successfully. It was thought that high scholastic grades and high scores on certain selected tests would have a positive relationship to successful work in the library.

A testing program was set up at the Iowa State College Library during the college year 1937-38. The purpose of this program was to discover the extent to which academic grades of student assistants and scores made on certain selected tests might be used to predict success in various types of college library work. The results reported here are based upon data obtained for 307 undergraduate student assistants¹ who worked in all departments of the college library during the college years 1937-38 through 1939-40. The predictive indices available for this group included the *American Council on Education Psychological Examination* scores, the *National Institute of Industrial Psychology Clerical Test* (American Edition) scores, and the grade-point averages for one quarter of college work. In addition scores on the *Bell Adjustment Inventory* were available for 69 assistants who were included in the group of 307.

All students take the *Psychological Examination* when they

¹The group was composed of 174 freshmen, 86 sophomores, 37 juniors and 10 seniors. Of this number, 71 were women and 236 men. One hundred and thirteen students (40 women and 73 men) had had some previous library experience at the time they took the Clerical Test while 194 (31 women and 163 men) had had no previous library experience. "Library experience" as used in the study may be defined as more than four weeks of work, usually on a part-time basis, in a library. An assistant who has worked four weeks or less was considered in the group "without library experience"

enter college. The grade reported in the study was the grade-point average² made by the student assistant for the quarter in which he took the *Clerical Test* at the library. The library rating was made at the end of the same quarter.

The criteria of success for student assistants in college library work were (1) ratings made by librarians who supervised the work of assistants and (2) records of student promotions within the library.

The graphic rating scale used was adapted from one described by Filer and O'Rourke (3) and by Symonds (4, 66-68), and instructions similar to those described by Symonds were given to each rater. The staff member best acquainted with the student's work and directly in charge of it rated the assistant. The ratings were scored by assigning numerical values to the five different divisions of the rating line, with a possible range from 0 to 4 on each division and a total range from 0 to 40 on the ten items.

Only one rating was used in the study since it was found that very often there was no second person equally competent to make the rating. As a measure of the reliability of the ratings used, a second rating was made for two smaller groups of assistants included in the group of 307. Twenty students in the Catalog Department were rated independently by a second rater, and twenty students were rated by the Assistant Loan Librarian as well as by the Loan Librarian. Reliability coefficients of .76 and .77 were obtained. A frequency distribution of the total scores made on the ratings for the 307 assistants was made, and the chi-squared test indicated no significant departure from normality.

To be promoted, an assistant must not only have had some experience in the library, but he must have also the ability to perform more difficult tasks than those assigned to him when he began his library work. Student assistants who prove to be accurate in their work and who have the necessary personal qualifications are eligible for promotion. The records of stu-

²See Iowa State College Catalog, 1939-40. p. 118.

PREDICTION OF SUCCESS IN COLLEGE LIBRARY WORK

dent assistants who were promoted were obtained from the pay rolls.

The statistical methods used in analyzing the data included a study of the significance of differences of means and of the significance of correlation coefficients, and the use of regression coefficients.

Relationship between the Predictive Variables and the Library Rating

Table 1-A shows means and standard deviations of the

TABLE 1—A
MEANS AND STANDARD DEVIATIONS ON FOUR VARIABLES
FOR TOTAL GROUP OF STUDENT ASSISTANTS

	No.	N.I.I.P. C.T.	Grades	A.C.E. P.E.	Library Rating
<i>Means</i>	307	73.6	23	98.2	24.3
<i>S. D.</i>	307	19.59	.69	22.09	6.27

TABLE 1—B
COMPARISON OF MEANS ON FOUR VARIABLES FOR GROUPS OF STUDENT
ASSISTANTS CLASSIFIED ACCORDING TO LIBRARY EXPERIENCE AND SEX

Group	No	N.I.I.P. C.T.	Grades	A.C.E. P.E.	Library Rating
Men with library experience	73	74.7	2.434	101.9	25.4
Men without library experience	163	69.8	2.274	94.1	22.9
Difference		4.9	.160	7.8*	2.5*
Women with library experience	40	83.2	2.435	107.8	26.5
Women without library experience	31	78.5	2.351	98.7	26.1
Difference		4.7	.084	9.1	.4
Women with library experience	40	83.2	2.435	107.8	26.5
Men with library experience	73	74.7	2.434	101.9	25.4
Difference		8.5*	.001	5.9	1.1
Women without library experience	31	78.5	2.351	98.7	26.1
Men without library experience	163	69.8	2.274	94.1	22.9
Difference		8.7*	.077	4.6	3.2*

(*Indicates that the difference is significant)

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

four variables for the total group of 307 student assistants. The comparison of means in Table 1-B indicates that although the women with library experience made higher scores on the four variables than the women without library experience, the mean differences were not significant. Men with library experience made higher scores on the four variables than men without library experience and the mean differences for the *Psychological Examination* and the Library Rating were found to be significant. Women with library experience made higher scores on tests and grades and were given a higher rating than men with library experience, but the mean difference was found to be significant for the *Clerical Test* only. Women without library experience made significantly higher scores on the *Clerical Test* and were given a significantly higher rating than men without library experience.

Table 2 shows the correlation coefficients for the three independent variables with the library rating for the total group and for the subgroups. The correlation coefficients are positive but not high.

TABLE 2
CORRELATION COEFFICIENTS OF THREE VARIABLES WITH LIBRARY RATING
FOR GROUPS OF STUDENT ASSISTANTS CLASSIFIED ACCORDING
TO SEX AND LIBRARY EXPERIENCE

Group	No in Group	First-order correlation coefficients			Multiple R
		N.I.P.C.T.	Grades	A.C.E.P.E.	
Women with library experience	40	.205	.256	.293	
Women without library experience	31	.249	.479*	.311	
Men with library experience	73	.487*	.250*	.112	
Men without library experience	163	.351*	.424*	.246*	
Total Group	307	.393*	.377*	.263*	.456*

(*Indicates that the correlation coefficient is significant)

PREDICTION OF SUCCESS IN COLLEGE LIBRARY WORK

TABLE 3

CORRELATION COEFFICIENTS OF FOUR VARIABLES WITH LIBRARY
RATING, AND INTERCORRELATIONS
(Sample Fall Quarter 1939, N = 69)

Variable	Psychological Exam.	Clerical Test	Adjustment Inventory	Grades	Library Rating
Psychological Exam		.698	-.020	.570	.417
Clerical Test			-.070	.501	.622
Adjustment Inventory				.0001	.030
Grades					.468

Table 3 shows correlation coefficients and intercorrelation coefficients for a group of 69 students who were given an additional test, the *Bell Inventory*, during the Fall Quarter 1939. In the hiring and selection of student assistants in college library work, the measurement of ability is of prime importance. However, assistants must be able not only to do the work assigned, but they must also be able to get along with people and to have certain characteristics such as co-operativeness and dependability. Bell (1, 102-104) and Tyler (5) have reported very low correlation coefficients between the *Adjustment Inventory* and the *Psychological Examination* and the *Inventory* and grades. Since the rating scale used in this study contained items such as co-operativeness, initiative, and dependability, it was thought that a higher correlation coefficient might be obtained with the library rating than with the measures of intelligence. However, the correlation coefficient of .03 between the *Adjustment Inventory* and the Library Rating is so low as to be of no value as a predictive device for the selection and hiring of assistants. Negative correlations were obtained between the *Adjustment Inventory* and the *Psychological Examination*, and between the *Adjustment Inventory* and the *Clerical Test*. Bernreuter (2) points out that this type of inventory should not be used in selecting individuals for jobs since the complete co-operation of the individual is essential and this complete co-operation is difficult to obtain in a program which involves selection for jobs. Further research is needed to discover some instrument which

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

will measure more satisfactorily the personal qualifications of the applicant. Very often it is possible to obtain through a carefully conducted interview an estimate of personal qualifications which would make for or prevent the success of an applicant

Standard regression coefficients for the total group and for the Fall Quarter 1939 group are shown in Table 4. These

TABLE 4

STANDARD REGRESSION COEFFICIENTS SHOWN FOR THREE PREDICTIVE VARIABLES FOR THE TOTAL GROUP AND THE FALL QUARTER 1939 GROUP

Group	No.	N.I.I.P.		A.C.E. P.E.
		C.T.	Grades	
Total	307	.305	.265	-.047
Fall Quarter 1939	69	.593	.244	-.135

standard regression coefficients indicate that the *Clerical Test* contributed most to the library rating with grades second and the *Psychological Examination* third for the total group and for the subgroup, Fall Quarter 1939.

Relationship between Predictive Variables and Promotions

While student assistants who make average scores on tests or average grades and are given an average rating may get along fairly well in the library, the average does not represent a satisfactory goal. There is constant need for students who can do better than average work, and devices which will help to indicate the more promising assistants at the beginning of their work are important. It may be seen from Table 5 that

TABLE 5

SIGNIFICANCE OF MEAN DIFFERENCES ON FOUR CRITERIA BETWEEN THE GROUP PROMOTED AND GROUP NOT PROMOTED

Group	No.	N.I.I.P.		A.C.E. P.E.	Library Rating
		C.T.	Grades		
Promoted	42	88.26	2.76	112.38	29.12
Not promoted	265	71.23	2.26	95.98	23.50
Diff. of Means		17.03	.50	16.40	5.62
S. D. of Diff. of Means		3.25	.11	3.77	1.04
Diff. of Means		5.3	4.5	4.3	5.4
S. D. of Diff. of Means					

PREDICTION OF SUCCESS IN COLLEGE LIBRARY WORK

the means of the group promoted are higher on all the variables than the means of the group not promoted. Not only are the means higher but the differences between means were found to be highly significant.

Obviously not all of the students promoted were equally capable. Some of the scores on tests and grades of certain assistants in the group promoted fall below the average of the group not promoted. By eliminating the lower 25 per cent of scores made on the *Clerical Test*, the *Psychological Examination*, and grades for the group promoted, all scores below the average of the group not promoted were eliminated. The score on each test and the grade which divide the upper 75 per cent of the group promoted from the lower 25 per cent of the group was taken as the critical score and critical grade. The critical score is 76 for the *Clerical Test* and 104 for the *Psychological Examination* (1938 edition) and the critical grade is 2.38.

Since the results show that mean differences between the group promoted and the group not promoted are significant, high scores on the tests used and high grades may be considered to be of value as predictive devices in the selection of student assistants for college library work. The critical scores determined for this group may be used as a guide in selecting assistants in the future who would have high promise of success.

REFERENCES

1. Bell, Hugh M. *The Theory and Practice of Personal Counseling, with Special Reference to the Adjustment Inventory*. (rev. ed.) Stanford University Press, 1939.
2. Bernreuter, R. C. "The Present Status of Personality Trait Tests," *Educational Record*, XXI (1940), Supp. 160-171.
3. Filer, H. A. and O'Rourke, L. J. "Progress in Civil Service Tests," *Journal of Personnel Research*, I (1923), 484-520.
4. Symonds, Percival Mallon. *Diagnosing Personality and Conduct*. New York: Century Publishing Co., 1931.
5. Tyler, Henry. "Evaluating the Bell Inventory," *Junior College Journal*, VI (1936), 353-357.

THE ADMINISTRATION OF GROUP TESTS

ERNEST M. LIGON

Union College

“**A**NYBODY with a strong voice who can read can give group tests.” Unfortunately this opinion is very widely held. Even a superficial consideration of the responsibility of a group test situation should quickly dispel such an idea. Actually, good group testing is much more difficult than individual testing. The most perfect tests available are as valueless without good examiners as the best surgical instruments without good surgeons.

Among the prerequisites of good group testing are: that all of the subjects understand the instructions, that they all work throughout the assigned time at their optimum level of achievement, that they are in no way helped, hindered, or distracted by one another, that they do not quit trying or omit any section of the test, that examiners give instructions adequately and in a stimulating, effective tone of voice—not a dull, bored monotone—and that proctors are observing every movement in the group, stimulating lagging souls, inhibiting wandering eyes, and detecting failure to follow instructions. Literally millions of group tests are administered every year. On scores derived from them an equal number of judgments are made affecting in some way, often extensively, the lives of those taking them.

An examiner giving an individual test can easily determine how his subject is reacting to the test problems. A group test examiner has that responsibility for as many subjects as are in the room, which may range from only a few to several hundred. Now that group tests are being called upon to play such a large role in our war effort, it behooves us more than ever

to make every effort to give them effectively, as well as to construct them carefully and score and interpret them accurately. Every man who, due to no fault of his own, makes a score on one of these tests which does not reflect his true capacity may thereby be put in the wrong place. It does not require proving that in so specialized an organization as the modern mechanized army, a very important part of its success depends on getting the right man in the right place.

This paper has been prepared on the basis of two types of evidence. In the first place, the author has had several years of experience in administering many different kinds of group tests as well as individual tests. During this period, it has also been necessary to train many students to do so. In the second place, although the literature contains very little either in periodicals or in books on measurement concerning this phase of measurement procedure, almost all group tests include in their manuals of procedure such instructions as seem desirable for their administration. A number of these have been examined and the principles included in them collected and organized into this paper. The ones used were selected simply because they were ones with the administration of which the author has had wide experience. It seems probable that they constitute a fairly representative sample.

I

The aim of a group test is to measure *differentially a group of homogeneous¹ individuals with respect to some simple or complex variable.*

Basic requirements, if scores are to be significant, presuppose that all subjects

- (1) give their optimum performance, and
- (2) do so for the full period of the allotted time.

¹By *homogeneous* is meant that group tests are always administered to a group selected because all of its members can be measured with respect to the variable or variables involved and by means of the test being used.

THE ADMINISTRATION OF GROUP TESTS

These, in turn, presuppose for the examiner

- (1) that he make perfectly clear to the subjects what they are to do and how they are to do it;
- (2) that he stimulate them to do their best;
- (3) that he and his proctors note and make adequate adjustments for individual deviations, such as mental confusion, indifference, impulsiveness, day-dreaming, making right responses in a wrong way, cheating, and the like, which might destroy the validity of test results.

II

The most common sources of error in group test administration and methods for controlling them follow.

(1) *Misunderstood instructions*

A very common misconception is that if printed instructions are read word for word, this is sufficient to hold conditions constant. The fact is that, unless the test itself consists of instructions, *conditions are held constant only when every subject starts to work with a complete and accurate understanding of what is expected of him.* Furthermore, the method of reading instructions is quite as important as their wording, so far as the subjects' ability to comprehend them is concerned. All instruction manuals ought to be marked as to emphasis and pauses, as well as carefully worded. Such a practice would add substantially to the accuracy of group tests.

The speed of reading instructions should be a function of the speed of comprehension of the subjects. The alert examiner, by watching his subjects, can know when his subjects understand what he has just said. Some instructions, therefore, can be read more rapidly than others, and with some groups of subjects more rapidly than with other groups. The time for reading should be a little more than adequate for all subjects.

Enunciation is important, considering that in most large groups, such as in the army, several "dialects" will probably be represented. All must understand.

The auditory acuity of the subjects must be considered. Instructions are more certain to be understood if the subject can read them silently as the examiner reads them aloud, thus emphasizing both visual and auditory cues. Conversely, however, the examiner must read them aloud. To ask subjects to read instructions silently without oral reading by the examiner almost always results in errors in reading and even failure to read all of them. Furthermore, as previously indicated, the emphases indicated by the reader help in understanding. Visual illustration of the mechanism of recording answers will help to avoid many clerical errors.

Delayed reaction instructions ought to be repeated near the time during the test when they are to be carried out. Instructions about what to do when a certain part of the test is reached are almost certain to be forgotten by some subjects unless they are reminded of them.

Proctors should check to be sure that all of the subjects do understand the instructions, by seeing, for example, how they answer the first two or three questions, usually simple ones which all ought to answer.

In some tests, ability to understand instructions is part of the test. If this is to be the case, there should be a great many different instructions separately scorable. Otherwise, failing to understand them may produce an undistributed minimum. Occasionally in screen tests this may be a desirable condition. It is also true, at the other extreme, that if instructions are so complete as to constitute answers to the test questions, there may be an undistributed maximum. It seems unwise, however, to include in the time limits of a test nonscalable sets of instructions.

Added afterthought directions given in the middle of a test by an examiner who has not adequately prepared his instructions beforehand constitute an important source of distraction.

The amount of practice necessary to make instructions clear to subjects will vary among various groups. It needs to

be adequate for the poorest of the subjects. One hundred per cent understanding is necessary if subjects are to be measured accurately.

(2) *Careless Errors*

Speed and accuracy are two distinct qualities. Subjects should know how much each is to be weighted in the scoring of a given test. For example, a score based on right minus wrong on a clerical aptitude test can be raised appreciably by working very rapidly even though the number of errors is thereby increased. However, most employers of clerical workers would rather have employees who get sixty correct answers out of sixty attempted than eighty-five right out of one hundred attempted, although the score of the latter is higher than the former.

Persuading subjects who finish before the time is up to reread the questions and check the answers is a help in avoiding careless errors. Proctors need to be alert to the possibility of subjects overlooking large sections of the test.

Whether or not subjects are to be encouraged to guess on multiple choice questions should be standardized. It is common procedure for examiners to warn the subjects that a wrong guess counts off more than an unanswered question. They neglect to add that a right guess counts more than an unanswered question.

Careless errors which are a result of the faults of the examiner need to be watched for. Timing errors are the most common of these. A full-face second hand is a necessary part of a group testing time-piece. Extra pencils need to be at hand, so that securing a new one from a proctor does not require a large amount of time. If subjects have two pencils to begin with, the possibility of this difficulty is decreased. Pens should never be used, since the inevitable corrections made by a subject become increasingly difficult or even impossible to interpret.

Correct filling in of the forms on the front of the usual test blank is difficult. These need to be kept at a minimum and

filled out systematically under instructions from the examiner. The date should be stated or written in a prominent place in the front of the room.

Distractions are to be eliminated as far as possible. A distraction is a subjective concept. It is whatever distracts the subject. Too quiet a room may often be more distracting than one which is noisy. Visual distractions are usually more important than auditory ones. People walking by where they can be seen, proctors too obvious observation, neighbors turning to later pages of the test too soon, some subjects leaving the room too early, and excessive materials on desks are a few of the more common distractions.

(3) *Low Motivation*

If group test scores are to be adequate measures of what they try to measure, they presuppose that the subjects do their best for the full time limit of the test. Unless the test measures motivation, maximum motivation is a prerequisite for accurate scores. There are at least three types of motivation which, when characteristic of group test subjects, tend to decrease the reliability of the results.

(a) *Sense of inadequacy.* Subjects often see that there are many problems on the test which are impossible for them to answer and infer therefrom that they are failing the test and so give up without trying. This is due to their experience with school examinations, in which they are expected to know all of the material asked for. A statement of the nature of mental tests will often remove much of this misconception, emphasizing especially that a good test must be long enough and sufficiently difficult that the best subject cannot make a perfect score. Then, too, a statement to the effect that too high a score is quite as bad as too low a score for a subject will help, emphasizing the fact that the accurate score is the only good one. This source of error is especially characteristic of achievement tests in fields in which the subject has had no formal training, such as mathematics and science, if that is the case. The subject often gives up without trying, whereas a

THE ADMINISTRATION OF GROUP TESTS

genuine effort would often produce astonishingly good scores, even if the subject has had few, if any, formal courses in these fields

(b) *Sense of indifference.* Many subjects may have the idea that the test is not important and that it does not make any difference what they make on it. There may be initial indifference, or a lack of enthusiasm for even starting a test, and there may be executive indifference, or a decrease in enthusiasm as the testing period progresses.

Overcoming *initial indifference* depends on (a) the attitudinal preparation of the subjects for the test, and (b) the attitude inspired by the test examiner in the beginning of the test.

As subjects are prepared for a test, they should be told the purpose of the test; what it tests and how one can know that it tests it. A brief statement is often very effective, pointing out that tests are constructed by experiment and not by arm-chair theory and should be criticized only when experimental data are available. This is especially important with highly intelligent subjects. A discussion of the nature of direct and indirect tests, with a clear statement of which type is being administered, is valuable. If a subject knows that he cannot predict the right answer and that he may only destroy his chances of getting a good score by endeavoring to do so, he is less likely to try to answer the questions in whatever way he thinks may get a good score instead of giving straightforward answers

The subject should also be informed as to the use to be made of the results. If they are confidential, to be given to no one except him or with his permission, his attitude is improved by assuring him of this fact. Honesty and frankness with the subjects is almost always an asset in getting good motivation.

Then, too, the attitude of the examiner is important. If he by his posture and tone of voice indicates that he is bored by the whole procedure, he will probably inspire this same

attitude in his subjects. If he stands erect and alert while reading directions, and speaks with a tone of enthusiasm in his voice which suggests that he thinks the test is interesting, the effect on his subjects will be appreciable. Just as in every individual test every subject is the "most important" subject, so in group tests, every group is the "most important" group. A good examiner never lets down.

The intensity of the examiner's voice needs to be controlled. However, subjective intensity is not always measured in decibels. It is a well known principle in public speaking that to lower the voice both in pitch and loudness is effective in getting attention. This probably is due to the fact that it is the very opposite of the common procedure. An incisive, firm, ringing signal "go" does much to produce good initial motivation. When loud-speakers are used, their value lies in the fact that more objective intensity can be gotten without greater apparent intensity on the part of the examiner. In any case, the attitude and posture of the subjects should be one of alert attention at the signal "go."

Overcoming executive indifference is even more difficult than overcoming initial indifference. The most important factor in this respect, aside from the personality traits of the subjects themselves, is the internal nature of the test. In young children, every test needs to be put on a game level, in order to elicit best efforts. If tests are of any considerable length, this also holds true even for those given to adult groups. Even when the subjects have the best intentions and the most complete awareness of the importance of the results, it is difficult not to let down with continuing boredom. Test construction and the organization of test batteries ought to be based on this inherent factor and provide for the inclusion of interest stimuli at frequent intervals. Test reliabilities would thus be improved. A spirit of competitiveness, if not overdone, is of value in executive motivation. It must be geared to the type of subject with whom the test is used, but is always stimulating when used wisely.

THE ADMINISTRATION OF GROUP TESTS

The attitudes of the examiner and proctors, even during the times when the subjects are writing and no instructions are being given, are still a factor. If they relax and slouch around in groups for non-test conversation, this will carry over to the subjects. Examiners and proctors who are alert during the whole testing period have an important role in the maintenance of motivation in their subjects.

Mental fatigue is largely a product of imagination. Some groups become tired after fifteen minutes. Others will continue at full speed for several hours. Subjects ought to be warned, in view of the importance of the results, about the shortsightedness of allowing fatigue and boredom to decrease the quality of their effort. They should be informed of the facts concerning the nature of mental fatigue and as to how long it is possible for the human mind to persist at a high level of effort and efficiency.

(4) *Mental confusion due to too great excitement*

It is possible, especially with some subjects, to get too great motivation as well as too little. The subjects need in the very beginning to be put at their ease, without a loss of desirable motivation. The methods employed by every good individual tester can with modification be applied in group testing. Thus, an individual tester adjusts the speed and intensity of his voice to the speed and intensity of his subject. If the subject is dawdling he can speed him up by a slight increase of these qualities in his own voice. If the subject is obviously too excited, he can quiet him down by the slower speed and lower intensity of his voice. This can be done also by the effective group tester.

Two common causes of overexcitement can be diminished by the good examiner. If the subjects worry about the tests a long time in advance, it may be wiser not to forewarn them at such long periods. No forewarning at all, on the other hand, might produce a complete mental disintegration in some subjects. Instructions which read "work rapidly," if given with too great fervor, may sometimes decrease the efficiency of cer-

tain subjects. Many people cannot work well under pressure of time or being hurried. The calmness and inspired confidence of the examiner can instill an alertness in his subjects without overexciting them.

When the instructions or the paraphernalia, such as machine-scoring equipment, are too complicated, plenty of time needs to be taken to familiarize the subjects with them to insure confidence in their use.

(5) *Dishonesty of Subjects*

Observing and copying one's neighbor's work is only one of the forms of cheating done by subjects on group tests. It is, furthermore, the easiest to control. It should simply be made impossible. Honor systems should never be used in group testing since relative scoring can be made invalid by even a few dishonest subjects. But there are other forms of dishonesty. Precheating is very common. Often people come to psychologists to "get a copy of all the tests to be had" so as to be ready for some coming group tests. Obviously, most tests cannot be prepared for. But subjects attempting to do so show by their attitude their unwillingness to give the most desirable type of co-operation. It stands to reason that to whatever extent a subject succeeds in this sort of effort, he destroys the value of the test to him as an accurate indication of his ability or aptitude. This type of cheating can best be eliminated by the process of urging upon the subjects the fact that a high score is a bad score unless it is accurate.

Getting information concerning the tests from individuals who have taken them is another dishonesty source of error. Subjects ought to be informed of the possible consequences which may arise from such action. If, for example, by this means one succeeds in getting into the air corps who would have been rejected if properly tested, and is killed in training, his informer can hardly be thought of as having done him a favor.

Such test procedure methods as involved in requiring "pencils up" between tests and "folding back booklets so that

only one page at a time is visible" are designed to eliminate errors belonging in this classification. Clear instructions as to whether it is permissible to proceed to another page or go back to a former page without further signal ought to be given both orally and on the printed page.

Many subjects get help from the examiners themselves. Some examiners indicate correct answers in their tone of voice. Others will answer individual questions which may give the questioner an undue advantage. Such questions, if answered at all, ought to be answered so that all can hear. An example of how the examiner's voice can help the subject is found in giving the digits test in the Binet. If the digits are grouped or read too rapidly the test is much more easily passed.

It is well for the examiner to have a correct attitude concerning the nature of his job. The job of a teacher is to help his students learn. The job of an examiner is to measure, not to teach.

(6) *Not working full time*

If time is a factor, a test should be so constructed that the best subject cannot finish within the prescribed time limit.

It is common procedure in achievement tests to give adequate time for even slow subjects to finish. This results in the best subjects finishing early. If they leave the room, this becomes a serious distraction to the slower ones. It might be desirable to include non-scorable questions to prevent this from happening. It is true that holding subjects after they have finished is usually not good for testing morale.

It is difficult on long tests for subjects not to let their minds wander from time to time. This is a factor for the proctors to deal with. If the proctors are alert, both by their presence and their active efforts, they can keep the subjects working consistently. The discussion of executive indifference is also related to this point. Subjects, of course, should know whether or not it is a timed test, and in the case of long tests should be warned at regular intervals as to the amount of time consumed or remaining for the test.

Another factor which enters into the timing problem of tests is that of mental set. When long tests covering entirely different types of material are given successively, a longer interval needs to elapse between them. This is not due primarily to a fatigue factor, but to the need for changing the mental set from one field to another. Teachers who have taught two different courses in successive hours will recognize the importance of this principle.

(7) *Size of group and group inter-distractions*

Good morale in a group is essential to maximum performance. How large a group can be before distracting factors enter in due to size, varies. One group of four hundred can be tested better than another group of fifty. The constitution of the group is a factor, as is the ability of the examiner. When the members of the group do not know each other, morale is more easily maintained than when they do, unless intergroup rivalries can be used as a motivation. Groups of older age levels are usually easier to control than younger groups. Groups competing with each other have better morale than those having no such sense of group solidarity. Telling a new freshman class or group of draftees that they are competing with preceding classes or groups is an incentive for group morale.

Once group morale is lost, it is very hard to regain. Let there be a few sighs, whistles, groans, shufflings of feet, low-intensity grumblings, or catcalls and the situation for good group testing is almost hopelessly lost. The leadership of the examiner and the alertness of the proctors will play a large part in this.

III

This paper has attempted to indicate the difficulties involved in the administration of group tests and to point out some methods for making them good measures of the variables involved. Every subject ought to leave the testing room feeling confident that he has done his best, and that the score

THE ADMINISTRATION OF GROUP TESTS

assigned to him will be representative of him, even if he has missed a large percentage of the questions. One does not pass or flunk tests any more than he passes or flunks measures of height and weight. It is the task of the examiner to make this clear to each subject and get from him a sample of his best performance. More thorough training of group testers and a larger sense of their responsibility among them will make the increasing use of group tests a far greater contribution to the problems of adjustment than if the common notion that "anybody can give group tests if he reads the printed instructions word for word" continues to be the prevalent one. It will be obvious that not all of these principles will be applicable to all group tests, but it should be equally obvious that administering any group test is difficult and, when well done, constitutes a highly skilled act

THE PURPOSE, ORIGIN, PLAN OF PROCEDURE, AND VALUES OF THE NATION-WIDE EVERY PUPIL SCHOLARSHIP TESTS*

H. E. SCHRAMMEL
Kansas State Teachers College

Purpose

IN the field of measurements and the objective testing movement, the *Nation-Wide Every Pupil Scholarship Test* is one of the major significant developments. Because of the far-reaching influence of these testing programs in this respect, it was felt that it would be worth while to recount before the membership of the National Association of Teachers of Educational Measurements the major details of their purposes, origin, methods of procedure, and values.

The purpose of the *Every Pupil Scholarship Test* is the promotion of scholarship. They are a valuable agency for stimulating scholastic endeavor on the part of the students. They stimulate good teaching as well as application to better learning. They vitalize education and make schools more worth-while in the lives of the students.

Origin

The *Nation-Wide Every Pupil Scholarship Tests* sponsored by the Bureau of Educational Measurements of the Kansas State Teachers College of Emporia had their origin twenty years ago in connection with the county and state Scholarship Contests sponsored by this college.

The first county contest in academic subjects, of which we find a record, was conducted by the Bureau of Educational Measurements in 1922 in Cloud County, Kansas. The first

*Paper read at meeting of February 24, 1942, in San Francisco.

State Scholarship Contest on record was conducted by the Emporia State College in 1923.

For a time the county contest movement was very popular, and the state contest movement also developed at a marked rate. The latter is still a popular event in Kansas. This spring the twentieth annual State Scholarship Contest will be conducted by the Emporia State College at thirty conveniently located centers of the state. Last spring over 3,500 students from approximately 200 high schools participated in this event.

In the county contests at first only a few of the best pupils participated from each school. Hence the suggestion was made that a plan be devised which would stress excellence in achievement of the entire class in a curricular field. Thus in the spring of 1924 two schools conducted a contest in one subject which involved a larger number of pupils from each school, each set of pupils taking the tests in their own school. The test papers were provided and scored by the Emporia State College. This was known as a dual contest. During the 1924-25 school year there was much demand for objective tests for use in similar inter-school contests in which every pupil in one or more specified subjects of each of the competing schools participated. Because of the increased demand for new tests for this purpose, a plan was devised for announcing in advance the subjects and dates for which tests would be made available for inter-school competition. During the first year that this plan was in operation, many schools used the tests for inter-school competition in which all the pupils of each school participated and the median score was used as the measure of comparison. A few schools, however, were not matched with any other schools for competition, but they desired to use the tests in order to be able to compare their results with the results of the other schools for the purpose of determining the relative excellence of their own classes. Hence norms were computed from all scores in each subject and provided to all the participating schools. Thus the Every Pupil Contest idea

PLAN OF PROCEDURE OF NATION-WIDE EVERY PUPIL TESTS

soon was superseded by the principle of a testing program in which schools voluntarily participate in order to obtain an objective measure of the attainment of pupils and classes. This is the plan that has been retained in the main, with the introduction from time to time of valuable perfections and improvements.

Plan of Procedure

At present the plan of procedure of the *Nation-Wide Every Pupil Tests* is as follows. The Bureau of Educational Measurements annually announces two dates for the testing programs. These come at the close of the first semester and near the middle of April. Bulletins are sent out giving the list of subjects for which new tests will be provided for each testing date. This year thirty-four new tests were provided for the testing program scheduled for January 8, and forty-four tests will be provided for the next testing program announced for April 8. Approximately 1,000 schools of the country obtain tests at mid-year and 1,500 for the end-of-year Test. About three-fourths of a million copies of the tests are used annually.

The Bureau secures competent volunteers to construct the tests. These consist usually of teachers in Kansas and elsewhere who are well trained in their respective curricular fields and who have also had some training in the field of measurement. The tests are edited at the Emporia State College by test construction and curricular specialists. The printing is done in the college print shop. Several dozen student assistants are employed in the office of the Bureau and in the print shop to handle the routine duties of typing, proofreading, filling and shipping orders, summarizing scores, computing percentile norms, invoicing, and keeping accounts.

As test orders are received from all parts of the country, norms are computed from the scores reported by the participating schools for each curricular field both for the whole

group and also separately for individual states from which a sufficient number of scores are reported to warrant it.

A summary bulletin of results is printed in compact form and furnished gratis to all participating schools within three weeks after the scheduled testing date.

For one of the recent Every Pupil Scholarship Testing Programs it was found that the process of computing the measures reported in the Summary Bulletin of Norms entailed the handling of 162,412 pupil and class scores, the construction of 405 frequency tables, and the calculation of 3,429 statistical measures. The norms computed are based on from several thousand to over ten thousand pupil scores for each of the various school subjects and grades for which the tests are provided.

Validity and Reliability

What method is used to insure that the tests possess adequate validity and reliability, the most important criteria for evaluating tests, is a question worthy of consideration at this point. While it is not claimed that the tests are fully standardized, they do compare favorably in these respects with the better standardized publications.

For insuring validity the following precautions are taken: First, as a rule the test builders are persons who teach classes in the curricular fields covered by the tests and who therefore have a good perspective of the content to be included. Second, content studies are made of textbooks and courses of study and the test items budgeted in accordance with the content distribution. Third, the editors consist of test construction specialists and supervisors and teachers of curricular fields. Fourth, cumulated studies of pupil responses on test items over a period of years are available and used. Fifth, cumulative criticisms from teachers who have used the tests over a period of years are available and utilized. Sixth, in fields where studies from previous editions of the tests are not available, preliminary editions are provided and tried out in representative classes.

PLAN OF PROCEDURE OF NATION-WIDE EVERY PUPIL TESTS

For insuring reliability, studies are made with preliminary editions and on tests provided over a period of years. In a field where tests are regularly provided, the degree of reliability may thus be predicted with a fair degree of accuracy.

The results are used extensively by teachers, principals, and pupils. Many expressions are annually received from schools as remote from the center of this movement as Montana, Florida, Texas, Maine, and California. Teachers are eager to learn how their classes rank in comparison with the classes of dozens of other schools in which the tests were administered on the same day during the current school term. Moreover, they want this information without much delay. It must be available promptly during the current school year to be of maximum value to them. Thus far we have been able to live up to the goal of mailing the results to the schools within three weeks from the day the tests are administered.

Pupils, too, are eager to note how they rank in comparison with the pupils in their own and other schools and they want this information before it becomes ancient history. By providing objective measures which can be simply and intelligently interpreted, pupils are motivated to work for greater excellence in achievement in the various curricular fields.

Many principals conserve the test results from year to year by filing the cumulative record of each pupil on a convenient card which has been provided. Because all scores are similarly interpreted, this provides a wealth of material for use in counseling and personnel work. Some schools also issue certificates of excellence to pupils whose scores receive a high percentile rank. In this manner excellence of achievement is further stressed and motivated.

Values Accruing from the Plan

The values accruing from the Every Pupil Scholarship Tests are manyfold. These may be roughly classified as pri-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

mary and as secondary values. The major primary values are the following:

- A. The plan of *Every Pupil Scholarship Tests* stimulates an intelligent interpretation of test results. For this purpose, percentile scores are provided for each subject and grade. Simple instructions are given for interpreting class median scores, as well as individual scores, into corresponding percentile scores. Because many different methods of interpreting standard test scores are resorted to, teachers are frequently at a loss in regard to the procedure in making correct interpretations. All too frequently valuable results from the use of standard tests are misinterpreted or not interpreted at all. Through our process of education in this respect over a period of years, many teachers and principals have exhibited that they have learned to make correct and meaningful interpretations of their results by the percentile score method and that they like the simplicity of this method.
- B. The plan motivates pupil and class effort because the results are objective and the interpretation is intelligible not only to teachers but can be presented graphically and intelligently to pupils.
- C. The plan motivates teachers in the construction and use of better home-made tests.
- D. The plan motivates teacher effort in better planning of instruction, finding of weaknesses of instruction, and so on.
- E. The plan motivates diagnosis of weaknesses and efficient remedial work in instruction.
- F. The plan challenges the teacher to set up outcomes objectives and to look for methods of determining the extent to which such outcomes are realized. Too frequently teachers are content in the assumption that valuable intangible outcomes accrue from their instruction, when in reality this may be far from the true conditions. By being consistently exposed to objective measurements, it is hoped that in time they will become skeptical of these assumptions and seek to evaluate the actual lasting values of their efforts.

PLAN OF PROCEDURE OF NATION-WIDE EVERY PUPIL TESTS

Among the secondary values, the following are a few of the more obvious:

- A. The plan aids the test builders to become more proficient in devising more efficient and valuable tests.
- B. On the Emporia State College campus the plan aids several dozen students annually who need employment to finance their college education.
- ~~62~~ C. For the Measurements classes on the campus, the plan provides an invaluable laboratory. All of these students are in some concrete measure exposed to test production, standardization, use, scoring, interpretation, and so on.
- D. The plan affords unusual opportunity for teaching students in Measurements classes and other employees the use of mechanical devices in handling statistical and other data. For example, the Bureau office contains hand and electric calculating machines, comptometer, clip boards, postal rate scales, postal wrapping device, and Dictaphones. The college print shop is equipped with linotype, rotary press, folding machine, stapling machine, and other equipment essential to a modern printing establishment. A large number of students receive first-hand experience in the operation of these devices in connection with their employment made possible by the *Nation-Wide Every Pupil Scholarship Tests*.
- E. The plan of the Every Pupil Testing Programs makes it possible to standardize more and better tests than would otherwise be possible. Where normally scores for norms would be difficult to obtain, and at considerable cost, a much larger sampling is possible for the norms and at practically no cost. This makes it possible to pass the advantages on to the patrons in terms of more and better up-to-date tests at an unusually low cost of production. During the writer's directorship of the Bureau of Educational Measurements, fifty-seven tests have been standardized. Most of these are published by the Emporia State College, but a few are published by some of the other leading test publishers.

- F. For many schools, participation in the *Nation-Wide Every Pupil Scholarship Tests* has furnished excellent material for local school publicity. In this manner taxpayers and patrons are made aware that their schools are not only seeking to excel in the so-called extra school work, but also in the regular curricular fields.
- G. Through use of these tests, capable pupils, who otherwise might be content to terminate their education upon completion of junior or senior high school, make a discovery of their own scholastic potentialities and are inspired to seek further development in college.

A TEST FOR SELECTING AND TRAINING INDUSTRIAL TYPISTS

CLIFFORD E. JURGENSEN
Kimberly-Clark Corporation

THE *Typing Ability Analysis* reported here was developed to assist in training typists for industrial positions, and to assist in hiring typists who can adequately fill such positions. It was developed after tests commonly used in high schools and business colleges had been found to be valueless in predicting typing success in stenographic and secretarial positions in Kimberly-Clark Corporation.

Analysis of reasons for the failure to predict job success by means of customary typing tests indicated that such tests do not emphasize sufficiently the major factors found in the industrial situation. Most typing tests provide an adequate measure of the mechanics involved in speed and accuracy when typing from printed copy. They fail to measure the mechanics of handling paper, placement of paper, use of tools, etc. They also fail to measure the non-mechanical aspects of the job which are the major factors in differentiating between successful and unsuccessful typists. Important non-mechanical aspects which should be measured include following instructions, noting and correcting errors, and the typing of diverse kinds of material rapidly, accurately, and in good form. Usual typing tests emphasize straight copying, and neglect the mosaic involved in a composite typing job.

Construction of the Test

Job analyses were inspected and conferences held with supervisors of industrial typists to determine the kinds of typing most often done, errors most often made, characteristics

which differentiate between successful and unsuccessful typists, etc. Supplementary data were secured by inspection of the corporation's files, and conferences with private secretaries and others in key typing positions. After analyzing and classifying the data obtained, files were inspected to obtain representative samples of the various kinds of typing. These samples were modified so they would be suitable for a test of typing ability, modification consisting primarily in the use of fictitious names and addresses. Standard typing procedures were substituted for those unique to the corporation concerned, in order that persons unfamiliar with procedures within the corporation would not be penalized unfairly.

After collecting and adapting twenty samples of different kinds of typing, instructions were prepared for each. Extreme care was taken that these instructions emphasize factors previously found important in differentiating between successful and unsuccessful typists. The preliminary test form of twenty items was administered to typists ranging from those known to be unsatisfactory to highly successful private secretaries. Tests were administered individually, and a record was kept of the time required to complete each test part, errors made in each part, comments on the test before it was scored, and comments of the typist after she had been told how her test results compared with those of other persons. This procedure was followed with a group of thirty typists half of whom were fully experienced high caliber stenographers or secretaries, and half of whom were inexperienced typists in the mailing department. Unanimous agreement among supervisors acquainted with each typist was required for inclusion in one of these two groups. Although the number of girls included in each group was small, the number was considered sufficiently large to warrant preliminary modification of the test. As a result of this tryout, the entire test was extensively revised, and seven of the twenty work samples were eliminated. The revised test was administered to a group similar to the one first used, the procedure and conditions of administration re-

A TEST FOR SELECTING AND TRAINING INDUSTRIAL TYPISTS

maining the same. Six additional work samples were eliminated on the basis of low validity coefficients or high intercorrelations with remaining parts. Accumulation of additional data has subsequently resulted in elimination or modification of other parts, and the present test consists of five carefully selected work samples. One of these is used as practice material so that the final score is based upon four selections. These parts are described later.

Considerable attention was given to developing a test which not only has a high validity coefficient, but which also appears valid to persons to whom the test is administered and to supervisors of such persons. It has been the author's experience that apparent validity is of equal importance with statistical validity. A test must have both if it is to be used successfully for industrial selection and training.

Development of Error Scores

The test was originally scored for errors in such a way that the test situation was as comparable as possible with actual work situations. Test results were examined from the viewpoint of whether or not similar work would be accepted by supervisors if submitted by an employed stenographer. Penalties for errors varied in direct proportion to the time required to make the work acceptable. No penalty was given for errors which did not affect the acceptability of the work, such as neat erasures. Errors of such a nature that the item would have to be retyped in order to be usable were penalized in proportion to the time required to type that item. This was subsequently modified so that the penalty was in proportion to the time required to retype the item inasmuch as it was found that the retyping time was not proportional to the original typing time. Errors that could be corrected so that material could be used in an actual work situation were penalized in proportion to the time required to make the necessary corrections.

Statistical analyses showed that the use of maximum penalties reduced the validity of the test by preventing differen-

tiation between poor typists. For example, in one test part, the testee must alphabetize the material, tabulate in three columns, make a carbon copy, etc. A testee was given the maximum penalty if she neglected to alphabetize the material; others who made all of the errors listed above would be given the same penalty as the first girl. Although the error scores for these girls would be identical, the quality of work on the test item concerned would be far from the same.

Originally no penalty was made for corrected errors if erasures were neatly made. The assumption was made that girls making erasures automatically penalized themselves by increasing the time required for completion of the test. Statistical analyses showed, however, that validity coefficients were increased by penalizing such corrected errors.

On the basis of errors made in the preliminary forms of the test, an item analysis was made of the seriousness of each error. This analysis showed that three classes of errors were sufficient for a total error score. The three classes were named: (1) corrected errors, (2) minor errors, and (3) gross errors. On the basis of the item analysis, the following three definitions were established to explain the three classes and to assist in the subsequent determination of the seriousness of errors found so infrequently that they could be given no statistical weight:

Corrected Errors are those which have been corrected by the typist (e.g., neat erasures). Each unit correction (whether it be a letter, word, or phrase) is counted as one corrected error.

Minor Errors are those which are correctible (e.g., misspelled words, strike-overs, etc.) or which detract from the form, arrangement, or neatness of the finished work to the extent that the material is acceptable for use but is below the desired standard.

Gross Errors are those which cannot be corrected unless the work is retyped (e.g., failure to make a carbon copy, failure to tabulate the material, etc.), those which are equivalent to two minor errors, or those which result in form, arrangement, or neatness below the minimum standard of acceptability.

A TEST FOR SELECTING AND TRAINING INDUSTRIAL TYPISTS

Error weights for gross, minor, and corrected errors were statistically determined by means of biserial validity coefficients for each type of error in a group of 63 employed typists, the criterion of success being grade of job successfully filled. Beta weights in raw score form were obtained through application of the Wherry-Doolittle test selection technique (5) using intercorrelations based on 250 applicants for typing positions.¹ In order to simplify scoring procedures, beta weights were rounded to the nearest whole number. The total error score is the sum of the corrected errors plus two times the sum of minor errors, plus four times the sum of gross errors. Data are summarized in Table 1.

TABLE 1
DEVELOPMENT OF ERROR WEIGHTS

Type of Error	Biserial Validity r	Beta Weight Z-Score Form	Raw Score Beta Weight	Rounded Weight
Gross	-.521	-.3421	-.3212	4
Minor	-.639	-.4700	-.1622	2
Corrected	-.430	-.3070	-.0915	1

Use of Combined Speed-Accuracy Score

In some cases it is desirable to interpret test scores from the two viewpoints of speed and accuracy. Usually, however, a combined score is preferable inasmuch as speed is worth little if not accompanied by accuracy, and accuracy is worth little if not accompanied by speed. Further, any typist can increase her speed at a sacrifice of accuracy or improve accuracy at a sacrifice of speed; therefore a combined score which is a function of both speed and accuracy will describe the typing performance better than either speed or accuracy alone.

Usual methods for combining scores (such as converting to standard scores or weighting by the reciprocal of the standard deviation) cannot be used with these data inasmuch

¹This procedure assumes that the larger group is comparable in all relevant respects with the smaller (criterion) group. The multiple correlation obtained from intercorrelations based on the expanded group may be larger or smaller than that obtained from intercorrelations limited to the criterion group. The use of an expanded group, however, has been found to increase test validity when the test is used with another group in a follow-up study (4).

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

as the distributions are not of the same shape and error distributions are far from symmetrical (3). Time distributions closely approximate normality and error distributions are greatly skewed toward the high error end. This is as expected, because one end of the distribution will be zero errors, whereas there is no limit at the other end of the distribution.

Error scores were transformed into "converted errors" by means of the best fitting curved line in accordance with Horst's method (3). The resultant distribution, based on 636 cases, was normal, and was expressed in terms of a mean and sigma equal to that of the time distribution. Total error scores are changed into converted error scores by means of Table 2. The time score can be added directly to the converted error score to obtain a combined score giving equal weight to speed and accuracy. Obviously, the time and converted error scores can be weighted in any other desired manner.

TABLE 2
TABLE FOR CHANGING TOTAL ERRORS INTO
CONVERTED ERROR SCORE

T.E.	C.E.S.	T.E.	C.E.S.	T.E.	C.E.S.	T.E.	C.E.S.
0	26	15	69	30	86	48-49	98
1	30	16	70	31	87	50-52	99
2	34	17	72	32	88	53-56	100
3	38	18	73	33	88	57-60	101
4	42	19	74	34	89	61-65	102
5	46	20	75	35	90	66-71	103
6	49	21	77	36	91	72-77	104
7	52	22	78	37	91	78-83	105
8	55	23	79	38	92	84-89	106
9	57	24	80	39	93	90-95	107
10	59	25	81	40	93	96-101	108
11	61	26	82	41	94	102-108	109
12	63	27	83	42-43	95	109-114	110
13	65	28	84	44-45	96	115-120	111
14	67	29	85	46-47	97	121-126	112

T.E.= total errors obtained by 4 x gross, plus 2 x minor, plus corrected errors.

C.E.S.= converted error score which can be combined directly with time score.

Nature of Test

The *Typing Ability Analysis*² consists of five parts, each being complete in itself. The first part is not scored, and consists of typing identification material such as name, address, and date. Part Two consists of approximately 150 words of a draft of part of an article to be typed. The work copy is in typed form, but contains thirty-five errors which are marked for correction. Each error is accompanied by the correct form which is to be used. Part Three requires the tabulating of seven lines in three columns, together with appropriate column headings, title, etc. The fourth part consists of a letter ninety words in length. The letter is written in longhand and contains ten changes also made in longhand. Part Five requires alphabetizing and tabulating fifteen lines of authors' names, book titles, and publication dates, together with typing column headings and title.

All test parts contain instructions such as, "make a carbon copy on yellow paper," "type the heading in capitals and underline it," "place your initials and the present date in the lower left-hand corner," etc. Failure to follow directions is penalized. In a few cases penalties are made for items not specifically mentioned in the instructions; for example, failure of the typist to place the date or her initials on the letter. Such penalties are made only for fundamental errors and failure to follow universal practice as taught in all typing classes and required of all industrial typists. Table 3 contains a list of all probable errors in each test part, classed according to whether they are scored as corrected, minor, or gross errors.

The *Typing Ability Analysis* is a work-limit test, each girl being permitted to complete the test. The shortest testing time in 636 cases was 26 minutes, and the longest time was 120 minutes. The average (mean) time required by industrial applicants is 61 minutes, and by high school seniors is 78 minutes. Although these average times are lengthy when compared with other typing tests, the increased time is warranted

²Published by Science Research Associates.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 3

CLASSIFICATION OF ERRORS

PART II—ROUGH DRAFT			PART IV—LITTE		
	Gross	Minor Corrected		Gross	Minor Corrected
Not on page 15	x		Not on letterhead	x	
No carbon copy	x		No carbon copy	x	
Carbon not on yellow paper	x		Carbon not on yellow paper	x	
Not double spaced	x		More than 4 1/2" after last line	x	
Omission of word or phrase	x		3 1/2" to 4 1/2" after last line	x	
Did not make indicated change	x	x	No date	x	
Strikeover	x	x	No initials	x	
Misspelled word	x	x	Omit "encl."	x	
Poor appearance	x	x	Less than 3 lines for signature	x	
Other error	x	x	Did not make indicated change	x	
Corrected errors		x	Strikeover	x	
			Misspelled word	x	
			Poor appearance	x	x
			Other error	x	x
			Corrected errors		x
PART III—TABULATION			PART V—ALPHABETIZING		
Not on page 13	x		Not on page 9	x	
No carbon copy	x		No carbon copy	x	
Carbon not on yellow paper	x		Carbon not on yellow paper	x	
No title	x		No title	x	
Title not in caps	x		Title not in caps	x	
Title not underlined	x		Title not underlined	x	
Headings not underlined	x		Column headings not underlined	x	
Not tabulated	x		Omit 2 or 3 column headings	x	
Omit "1941" and/or "1942"	x		Omit 1 column heading	x	
Omit "Type of Paper"	x		Not tabulated	x	
Columns in wrong order	x		Less than 3 spaces between		x
3 or more figures out of alignment	x		columns		
1 or 2 figures out of alignment	x		3 or more items out of alignment	x	
Omit word "total"	x		1 or 2 items out of alignment	x	
No line above "total"	x		Initials precede names	x	
No line below "total"	x		Line in wrong order (max 2 gross)	x	
"Total" lines not extended	x		Line omitted	x	
No initials	x		Incorrect publication date	x	
No date	x		Incorrect book title	x	
Strikeover	x		Strikeover	x	
Misspelled word	x		Misspelled word	x	
Incorrect figure	x		More than 5 punctuation errors	x	
Poor appearance	x	x	1-5 errors in punctuation	x	
Other error	x	x	Poor appearance	x	x
Corrected errors		x	Other error	x	x
			Corrected errors		x

by the high validity of the test. Inasmuch as the test administrator does no work after starting the test except to record the finishing time of the testee, the time required is of little practical importance to the administrator. The time element becomes important only if the typewriter being used cannot be spared for the required time, or if the testee objects to the time required. It has been the experience of the author that the practical appearance of the test tends to eliminate objections of testees to the time required.

In some cases it may be desirable to administer the test with a time limit, particularly when all that is desired is a yes or no decision as to whether or not an applicant should be hired. The time limit should be determined in such cases by deciding on the lowest percentile in terms of combined score which will be acceptable. Assuming no errors, the converted error score for no errors (26) should be deducted from the combined score representing the previously selected percentile. The resultant figure will give the time limit to be allowed. Or, if the administrator prefers, the time limit can be called when a desired percentage of the applicants have completed the test, all applicants who have failed to complete the test being eliminated from consideration. Use of the test with a time limit makes it impossible to secure a speed, accuracy, or combined score. Results therefore can not be used with maximum effectiveness for guidance or training.

Directions for Administering the Test

The *Typing Ability Analysis* is practically self-administering, and may be given either individually or in groups. In addition to a typewriter, each person taking the test should have the following materials: eraser, erasing shield, pencil, a sheet of carbon paper, 4 sheets of yellow paper for carbon copies, and a test booklet.

Before starting the test, each testee is permitted sufficient time to become familiar with the typewriter being used. Test booklets are issued with the instructions, "Read the instruc-

tions on the first page of the test. Do not turn the page or start the test until told to do so." Instructions appearing on the first page of the test booklet are as follows:

This test measures ability to do the kind of typing that is required in business and industry.

Test results will be judged by usual office standards and will be rated on the basis of accuracy, speed, and form. Errors will be penalized in proportion to their seriousness, least for errors which have been corrected, more for errors which could have been corrected, and most for errors which would require retyping of the part in which they occur.

Errors may be corrected by erasure. Do not re-type any part unless absolutely necessary, and in such case use the back side of the same sheet of paper.

Work as rapidly as possible, but do the kind of work desired by an employer.

After the instructions have been read, the examiner gives the signal to start the test. The exact time that each person starts and finishes the test is recorded. When the tests are completed, the examiner makes sure that the items are in correct order, and then staples all parts together. The time of starting and finishing is recorded on the first page of the test.

Scoring

The time (speed) score is the number of minutes required to complete the test.

The error score is based on three types of errors: (1) corrected errors, (2) minor errors, and (3) gross errors, an explanation of which was given earlier.

Each test part is proofread carefully and compared with the instructions. Errors are marked on the test by encircling them with a red pencil, and recorded on the rating sheet. The rating sheet contains a list of all errors commonly made, but is not a complete list of all possible types of errors. When unlisted errors are made, the definitions for errors as given previously should determine whether they are gross or minor.

A single error is not penalized twice. For example, figures out of alignment are penalized as such. An additional penalty is not given for poor appearance. If in copying the rough draft, the typist spells screen as "screne" a penalty is given for either misspelled word or failure to make indicated change. The error is not penalized in both ways.

When penalizing results for "poor appearance," the quality of paper is taken into consideration, inasmuch as neat erasures are almost impossible on some types of paper, but are easily made on other types.

Norms

As has previously been mentioned, a combined score will generally be more valuable than separate speed and accuracy scores. Correlational and regression equation techniques showed that for the typing jobs considered in this study, speed and accuracy were approximately equal in importance. Combined scores of maximum efficiency should be obtained by multiple correlations and regression equations based on test results of typists hired for each company using the test. Norms given here for three different weightings of speed and accuracy, however, will be adequate approximations for most jobs. The most suitable of the three will generally be the SA score which gives equal weight to speed and accuracy. It is obtained by adding the time score to the converted error score. The 2SA score (two times the speed score plus the converted error score) weights speed and accuracy in the ratio 2:1 and is suitable for jobs that require fast speed and where accuracy is comparatively unimportant (as might be the case for a rough draft copy typist). The S2A score (speed score plus two times the converted error score) weights speed and accuracy in the ratio 1:2 and is suitable for jobs placing a premium on high accuracy and where speed is comparatively unimportant (as in the case of some typists of legal documents).

Industrial and educational norms are given in Table 4. Industrial norms are based on 381 applicants for typing posi-

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

tions. Educational norms are based on 255 high school seniors who were given the test from one to two months previous to being graduated and who were in the advanced (second year) typing class.

TABLE 4
NORMS

Scale	Stan Score	Total 4, 31 111 Applicants Combined					Educational 255 H.S. Seniors Combined				
		Time	Errors*	SA	2SA	S2A	Time	Errors*	SA	2SA	S2A
100	3.00	16		54	78	76	32	0	76	116	104
99	2.33	26	0	71	103	103	41	1	91	139	128
98	2.05	30	1	77	113	114	46	2	97	148	137
95	1.64	37	3	88	128	130	51	4	106	162	152
90	1.28	42	5	96	142	144	57	5	114	174	164
80	.84	49	7	107	158	161	63	8	123	188	179
70	.52	54	9	115	170	174	67	10	130	199	191
60	.25	58	12	122	180	185	71	12	136	208	200
50	.00	61	14	128	189	194	75	14	142	217	209
40	-.25	65	16	134	199	204	78	16	147	225	217
30	-.52	69	20	141	209	215	82	20	153	234	227
20	-.84	74	25	149	221	228	87	24	160	245	238
10	-1.28	81	32	159	237	245	93	31	170	260	253
5	-1.64	86	41	168	250	259	98	38	177	272	266
2	-2.05	92	58	178	266	275	101	49	186	285	280
1	-2.33	97	81	185	276	286	108	66	192	295	290
M		61.41	17.20	127.96	189.37	194.50	78.78	17.14	141.76	216.64	208.63
S.D.		15.14	13.60	24.63	37.26	39.37	14.33	11.08	21.77	33.53	34.81

*Percentiles and standard scores computed on basis of converted error scores. For convenience, errors reported in this table are total error scores.

Norms are based on standard scores for selected percentile points. Analysis of data by means of the *Otis Normal Percentile Chart*³ showed marked linearity (normality) of all distributions. Standard scores can consequently be used to determine percentile points. The converted error score was used for computing error norms, though for convenience the errors reported in Table 4 are expressed in terms of total errors rather than converted error scores.

Correlations Between Speed and Accuracy

Correlations between speed (minutes) and accuracy (converted error scores) are all low. Summarized data are given in Table 5.

³Published by World Book Company

A TEST FOR SELECTING AND TRAINING INDUSTRIAL TYPISTS

TABLE 5
CORRELATIONS BETWEEN SPEED AND ACCURACY

Group	N	r	Standard error
All cases	636	+ .137	± .04
Industrial Applicants	193	+ .213	± .07
Civil Service Applicants	188	+ .200	± .07
High School Seniors	255	+ .077	± .06

Validity

Validity is based on 67 employed typists in an industrial population. Typists were dichotomized on the basis of grade of job successfully filled, and validity determined by means of biserial coefficients. The *p* group consisted of 28 girls employed in Kimberly-Clark Corporation's home office or in positions of similar caliber in various mills of the corporation. The *q* group was composed of 39 girls employed in mill offices and mailing departments of the same corporation. All girls were engaged in work which was primarily typing.

Guilford (2) has pointed out that: "a biserial *r* should not be computed unless the graduated series of measurements is reasonably well normally distributed and unless *N* is relatively large—preferably when *N* is greater than 50. Another important condition is that the cases be not too unevenly divided between the two distributions." These data fulfilled the above requirements reasonably well. The *N* of 67 was divided into *p* and *q* groups containing 42% and 58% of the cases. Pearson's chi-squared test of goodness of fit gave a *P* of .748 for a combined score based on equal weighting of speed and accuracy. Culler (1) classes this as an "excellent" fit. Validity coefficients are given in Table 6⁴

⁴It may be pointed out that an assumption underlying the derivation of biserial *r* is that the dichotomized trait is in reality continuous and normally distributed. If this condition does not hold, the size of biserial *r* may be appreciably affected; the value of *r* indicating perfect relationship may be considerably greater than unity and obtained *r*'s greater than would otherwise be obtained. It is entirely possible that this assumption was not fulfilled with these data, although the magnitude of the effect cannot be measured due to lack of methods which can be used to demonstrate normality of criteria in cases such as this. An attempt was made to approach normality so far as possible by including all grades of typists ranging from those in beginning typing jobs to those in the highest grade typing jobs of Kimberly-Clark Corporation.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

TABLE 6
VALIDITY COEFFICIENTS
(N = 67 Employed Industrial Typists)

Score	Validity r	Standard Error
Combined (SA)	.957	$\pm .04$
Time	.796	$\pm .08$
Converted errors	.711	$\pm .09$
Raw errors	.659	$\pm .10$
Gross errors	.456	$\pm .13$
Minor errors	.555	$\pm .12$
Corrected errors	.334	$\pm .14$

Additional evidence of validity was secured by comparing the differences in combined (SA) scores between various groups. Differences are summarized in Table 7. Critical

TABLE 7
COMPARISON OF GROUPS TO DETERMINE VALIDITY

Group	N	M	S.D.
A. Ind. typists—high job classification	28	91.89	14.04
B. Ind. typists—low job classification	39	135.05	20.85
C. Ind. typists—released, inadequate	10	178.50	21.71
D. Civil Service applicants—Jr. typists	63	122.00	20.71
E. Civil Service applicants—Asst. typists	125	130.22	19.57
Groups Compared	Critical Ratio	Significance	
A and B	9.97	.999	
B and C	5.18	.999	
D and E	2.60	.995	

ratios were computed by dividing the difference between the means by the standard error of the difference. The standard error of the mean was computed by formulae suitable for small samples (2), as follows:

$$\sigma_m = \frac{\sigma}{\sqrt{N-1}} \text{ when } N \geq 20;$$

$$\sigma_m = \frac{\sigma}{\sqrt{N-2}} \text{ when } 10 < N < 20.$$

All critical ratios are significant at the 1% level, thereby giving additional evidence of test validity by differentiating between groups which logically should show differences in ability.

Reliability

No adequate situation has yet been found in which to secure accurate reliability coefficients. Split-half methods are inapplicable inasmuch as test parts were selected on the basis of low intercorrelations as well as high validity coefficients. Two equivalent forms have been constructed; however, their use will usually be inadequate for reliability coefficients because of the effect of practice or disuse between testing periods. The same is true of repeated administration of one test form.

Table 8 presents reliability data obtained by administering two test forms to 63 high school seniors in their second-year typing class. The first test administration was in April and the second in May. Reliability is probably higher than indicated, inasmuch as practice between the two test administrations varied considerably; e g., some girls missed many typing classes because of commencement activities whereas others received considerable extra practice because of typing material for the school annual. Reliability coefficients quoted here are thus lowered by irrelevant influences of speed of learning and opportunity to learn. In spite of the unfavorable conditions under which reliability was secured, results nevertheless indicate reasonable reliability.

TABLE 8
RELIABILITY AND PROBABLE ERRORS
(N = 63 High School Seniors)

	Reliability		Probable Error of	
	Coefficient	Index	Obtained Score	True Score
Time	.768	.876	4.66	4.08
Converted Errors	.720	.848	5.47	4.64
SA Score	.832	.912	6.02	5.49
2SA Score	.846	.920	9.23	8.49
S2A Score	.800	.894	10.50	9.39

Interpretation and Use of Scores

The *Typing Ability Analysis* is now being used for several purposes and in several types of situations. The purpose for

its use in any particular situation determines the way in which results should be interpreted and utilized

Various companies are using the analysis for selecting typists for specific openings in stenographic and secretarial work. The test obviously should not be used as the sole means of selecting typists. Many factors must be taken into consideration, and this analysis measures only one group of such factors. The test should be used as a supplement to other selection methods and procedures, and test results should be interpreted in light of all other pertinent factors. Most companies using the test for employee selection do not discuss test results with applicants, although some believe that the time required for such discussion is warranted on the basis of improved public relations.

When the analysis is used for vocational selection, the industrial norms will usually apply because the employment manager will be interested most in knowing how a given applicant compares with other applicants. In the case of a recent high school graduate without any job experience, the employment manager may also wish to know how the applicant compares with high school seniors. Thus he is able to estimate not only how qualified she is at the present time, but also how satisfactory she is apt to be after securing typing experience.

A second major use of the analysis is in training currently employed typists. Results are discussed with the girl concerned and she is told whether her speed and accuracy are acceptable, the type (or types) of error she is prone to make, and other shortcomings which should be corrected in order that her work may be improved.

A third use of the analysis has been in the upgrading of industrial employees. Results have been used in the same way as in training in order to help typists prepare themselves for higher-caliber typing jobs, or to enable girls working in the plant on production jobs to fit themselves for typing jobs in the office.

A TEST FOR SELECTING AND TRAINING INDUSTRIAL TYPISTS

High schools are using the analysis for vocational guidance. Such usage generally includes suggestions for improving typing ability as well as recommendations regarding types of jobs which can be filled successfully. High school teachers will usually be interested in using the educational norms, although it may also be of value to determine how a particular student who will soon be a job applicant compares with other job applicants. Although high school teachers sometimes believe that such comparison is unfair due to lack of experience on the part of high school students, it must be remembered that most industrial personnel men are more interested in hiring an applicant with good typing ability than in hiring a promising high school student who compares favorably with other students but who cannot compete successfully with other job applicants. This is particularly true during depression periods when jobs are scarce and applicants are numerous.

The *Typing Ability Analysis* can also be used to compare the ability of various typing classes, efficiency of different teachers, rate of progress, etc., in all situations where typing ability is defined as those factors which differentiate between successful and unsuccessful typists in the industrial situation.

REFERENCES

1. Culler, E. "Studies in Psychometric Theory," *Journal of Experimental Psychology*, IX (1926), 169-194.
2. Guilford, J. P. *Psychometric Methods* New York: McGraw-Hill, 1936, 51-52, 351.
3. Horst, Paul. "Obtaining Comparable Scores from Distributions of Dissimilar Shape," *Journal of the American Statistical Association*, XXVI (1931), 455-460.
4. Jurgensen, C. E. "Extension of the Minnesota Rate of Manipulation Test," *Journal of Applied Psychology*, (1942) In press.
5. Stead, W. H., Shartle, C. L., et al. *Occupational Counseling Techniques*. New York: American Book Co, 1940, 245-252.

MEASUREMENT ABSTRACTS*

Bryan, Alice I. and Wilke, Walter H. "Audience Tendencies in Rating Public Speakers." *Journal of Applied Psychology*, XXVI (1942), 371-381.

Using the Bryan-Wilke Scale for rating public speeches, the authors studied a variety of audiences with a number of factors that are associated with audience ratings, such as factors related to analytical ability of audience, time of rating, effect of age of raters, influence of sex of raters, and intelligence and personality of speaker. *Louise T Grossnickle.*

Carter, Harold D. "How Reliable are the Common Measures of Difficulty and Validity of Objective Test Items?" *Journal of Psychology*, XIII (1942), 31-38

Two hundred college students, mostly juniors, were given an objective test, consisting of 80 items, of which 30 were true-false, 30 multiple choice, and 20 of the completion variety. The purpose was to ascertain the reliability of measures of item difficulty and of item validity by means of different subgroups. The author found that a measure of difficulty of test items, based upon a representative sampling, yielded a higher reliability coefficient than that obtained from the ordinary method of using good and poor students. *K. S. Yum.*

DuBois, Philip H. "A Note on the Computation of Biserial r in Item Validation." *Psychometrika*, VII (1942), 143-146.

A method of computing biserial coefficients of correlation through the use of punch card tabulating equipment is presented. Each item is assigned a separate column and successes

*Edited by Forrest A Kingsbury.

are punched 1. By arranging the cards on the criterion variable and obtaining progressive sums on several columns simultaneously, it is possible to obtain data for several correlations in one run of the cards through the machine. (Courtesy *Psychometrika*.)

Engelhart, Max D. "Unique Types of Achievement Test Exercises." *Psychometrika*, VII (1942), 103-115.

In this article are presented a number of unusual achievement test exercises of both the essay and the objective types. These exercises may suggest to others engaged in the construction of achievement tests certain forms which they may find useful either as models or as points of departure in the invention of new forms. The article also calls attention to certain problems which must be solved if achievement testing is to have a sound, scientific basis. (Courtesy *Psychometrika*.)

Estes, Stanley G. "A Study of Five Tests of 'Spatial' Ability." *Journal of Psychology*, XIII (1942), 265-271.

The object of the study was to determine the extent to which each of five tests, all of them requiring response to spatial relationships, were related to each other and to achievement in descriptive geometry, a subject where the ability under consideration was of basic importance. The correlations of four of these tests with descriptive geometry were all reliably greater than zero and did not differ significantly from each other. Therefore, the author concluded that the tests, with the exclusion of the Crawford Structural Visualization Test, were equally valid with the criterion he used. K. S. Yum.

Ferguson, Leonard W. and Lawrence, Warren R. "An Appraisal of the Validity of the Factor Loadings Employed in the Construction of the Primary Social Attitude Scales." *Psychometrika*, VII (1942), 135-138.

In this article the authors examine the effect of including alternate test forms in a factor matrix upon the validity of the resultant factor loadings, finding that in this particular instance the effect is negligible. Comparisons of the factor loadings derived from matrices in which only one of the alternate test forms is included with those in which both forms are included reveal practically no difference in the magnitude of either the original or rotated factor loadings, or in that of the computed communalities. (Courtesy *Psychometrika*.)

Ghiselli, Edwin E. "Estimating the Minimal Reliability of a Total Test from the Intercorrelations Among, and the Standard Deviations of, the Component Parts." *Journal of Applied Psychology*, XXVI (1942), 332-337.

Due to the nature of a test, two equivalent parts are not available for estimating its reliability. However, if, in all of the parts, sigmas are equal and the intercorrelations are equal, then the Spearman-Brown correction formula for any length can easily be derived from a general formula for the reliability coefficient of the total test. Since $r_{11} = r_{12}^2$ will be a minimum estimate of r_{11} , it is possible to obtain a minimal reliability coefficient of the total test. *K. S. Yum.*

Kelley, Truman L. "The Reliability Coefficient." *Psychometrika*, VII (1942), 75-83.

The reliability coefficient is unlike other measures of correlation in that it is a quantitative statement of an act of judgment—usually the test-maker's—that the things correlated are similar measures. Attempts to divorce it from this act of judgment are misdirected, just as would be an attempt to eliminate judgment of sameness of function of items when a test is originally drawn up. A "coefficient of cohesion," entirely devoid of judgment, measuring the singleness of test

function is proposed as an essential datum with reference to a test, but not as a substitute for the similar-form reliability coefficient. (Courtesy *Psychometrika*.)

Kuhlmann, F. and Odoroff, M. E. "Verification of the Heinis Mental Growth Curve on Results with the Stanford-Binet Tests." *Journal of Psychology*, XIII (1942), 355-364.

The usefulness of the I.Q. depends upon its constancy for the bright child and the dull child as well as for the typical child. However, this assumption is not warranted for all levels of intelligence. The late Dr. Kuhlmann preferred the index based on the Heinis mental growth curve because he believed in its superiority over the I.Q. for predictive purposes. This particular study, based on a large number of cases, shows that the average Stanford-Binet I.Q. of a group of special class pupils drops approximately 1.5 points per year or a total of 15 points between the ages of 6 and 16, while the average Heinis "personal constant," which Kuhlmann has renamed the "per cent of average" score, for the same cases shows no tendency to increase or decrease. *Louise T. Grossnickle*.

Moffie, Dannie J. "A Non-verbal Approach to the Thurstone Primary Mental Abilities." *Journal of General Psychology*, XXVII (1942), 35-61.

An attempt was made to measure five of the Primary Mental Abilities—perceptual speed, space, inductive reasoning, deductive reasoning, and memory—by means of performance tests. The author was successful in finding non-verbal measures of the space, reasoning, and perceptual speed factors. Tests for inductive and deductive reasoning were found to be measures of one reasoning factor. *Robert L. Cramer*.

Munroe, Ruth L. "An Experiment in Large Scale Testing by a Modification of the Rorschach Method." *Journal of Psychology*, XIII (1942), 229-263.

This technique of the Inspection Diagnosis consists essentially of a systematic review of each protocol with special attention to twenty-four items known to be of significance in Rorschach diagnosis. The results show some very striking correspondence between the Rorschach ratings and three separate lines of validation material, and suggest a strong probability that the Rorschach Inspection Diagnosis is a valid and useful technique for large scale testing. *Louise T. Grossmuckle.*

Powell, Norman J. and Levine, Harold. "Reliability of the Civil Service Oral Examination." *American Journal of Psychology*, LV (1942), 385-393.

Ninety-nine applicants who had passed a written examination for the position of Junior Psychologist were interviewed and rated in the conventional manner by two panels acting with varying degrees of independence. Considerable differences in ratings were found in all cases. *Robert L. Cramer.*

Stagner, Ross and Katzoff, E. T. "Fascist Attitudes: Factor Analysis of Item Correlations." *Journal of Social Psychology*, XVI (1942), 3-9.

Eighteen statements reflecting Fascist thought were presented to one hundred college students to be checked according to agreement or disagreement. A centroid factor analysis of the correlations between items showed three factors to be present: concern over protection of property rights, lack of sympathy for the unfortunate, and an aggressive nationalism. *Robert L. Cramer.*

Taylor, William S. "Partialling out Sums of Squares and Products in Calculating Correlations with Non-homogeneous Data." *British Journal of Psychology*, XXXII (1942), 318-323.

If the population tested is homogeneous, a coefficient of correlation calculated directly from the deviations of individual scores about the grand mean will give a reliable indication of the correlation between the scores of the individuals tested. Where the population is not homogeneous, group differences may be significant. The correlation desired is that free from the influence of group differences. In this latter case, it is necessary to partial out the sums of squares and sums of products of deviations from the mean, using only those attributable to the deviations "within groups." *K. S. Yum.*

Thomson, Godfrey H. "Following up Individual Items in a Group Intelligence Test." *British Journal of Psychology*, XXXII (1942), 310-317.

The article describes the technique used for item selection in the construction of Moray House Test 2+, a group intelligence test, and presents a later follow-up study of the test items in their discriminating function. The research reveals that the predictive power of the various test items differs with different levels of educational achievement. The items that predict well in the secondary school are not necessarily the best indication of their power to discriminate the potential secondary school pupils from those not suitable. *K. S. Yum.*

Thorndike, Robert L. "Regression Fallacies in the Matched Groups Experiment." *Psychometrika*, VII (1942), 85-102.

This paper is concerned particularly with certain regression effects which appear whenever matched groups are drawn from populations which differ with regard to the characteristics

being studied. It is shown that regression will produce systematic differences between these groups on measures other than those upon which they were specifically matched. The size and direction of these differences depend upon the differences between the parent populations both in the matching and in the experimental variables and upon the correlation between the matching and experimental variables. Formulas are presented for estimating the expected regression effect. Several alternative procedures are suggested for avoiding the erroneous conclusions which the regression effect is likely to suggest (Courtesy *Psychometrika*.)

Tinkelman, Sherman. "Civil Service Test Item Preparation: A Case Study." *Public Personnel Quarterly*, III (1942), 3-74.

The author traces the evaluation of test items to be used in a civil service examination, discussing source material, validity, public relations impact, and revision of the items. *Robert L. Cramer*.

Wolfe, Dael L. "Factor Analysis in the Study of Personality." *Journal of Abnormal and Social Psychology*, XXXVII (1942), 393-397.

A review of the previous studies in this field singled out seven factors, each of which had appeared in three or more studies. These were will, cleverness, shyness, self-confidence, fluency, depression, and hypersensitivity. The author noted two important characteristics of these personality factors. They will sometimes duplicate each other or sometimes cut across. Chief emphasis was placed upon the statement that factor analysis provides a powerful analytic tool for isolating the important variables of human personality and that the results thus obtained depend on the evaluation by clinicians and experimentalists. *K. S. Yum*.

Yum, K. S. "Student Preferences in Divisional Studies and Their Preferential Activities." *Journal of Psychology*, XIII (1942), 193-200.

The Kuder Preference Record was given to 193 college students for a study of their preferential interests in the seven major types, namely, scientific, computational, musical, artistic, literary, social service, and persuasive activities. The author found that the comparison of the mean profiles of the students in the physical, biological, and social sciences as well as the comparison of the mean profiles of men and women were significantly and consistently different on some of the major types of preferences. The correlation coefficients between the preference scores and academic achievement were negligible except in the case of the literary activities for the entire group and also for the group of men, and the computational activities for the group of women. *Louise T. Grossnickle.*

